Nanopore sequencing of ALIEN DNA

Christopher Alex Thomas

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Jens H. Gundlach, Chair

Paul A. Wiggins

Andrew H. Laszlo

Program Authorized to Offer Degree:
Department of Physics

University of Washington

Abstract

Nanopore sequencing of ALIEN DNA

Christopher Alex Thomas

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

Nucleic acids in the forms of DNA and RNA carry the genetic information of all known living organisms, enabled by an aperiodic crystalline structure in which four nucleobases, adenine, cytosine, guanine, and thymine (A, C, G and T) are arranged sequentially along a repeating sugar-phosphate backbone. How and why life came to adopt these specific nucleobases for genetic information storage is a mystery, and there are strong suspicions that other nucleobases could form robust genetic information storage systems. A variety of artificial genetic systems have been synthesized, retaining the backbone of DNA but incorporating non-standard nucleobases in addition to (or replacing) A,C,G,T. These artificial systems have applications in molecular diagnostics and targeted medicine, but are largely incompatible with modern techniques of nucleic acid sequencing. By tackling the sequencing problem, we can help realize the full potential of these applications by increasing the effectiveness of artificial nucleic acid quality control and sample reproducibility.

For my dissertation in the UW nanopore biophysics laboratory, I developed nanopore sequencing technology of an expanded eight-letter "*hachimoji*" DNA alphabet containing A,C,G,T as well as four additional hydrogen-bonding

nucleobases **B**, **P**, **S**, and **Z**. In nanopore nucleic acid sequencing, a single nucleic acid molecule is pulled through a nanoscale pore by an applied electric field while a bound motor enzyme controls its motion. The the bases of the molecule block the electric current through the pore with magnitudes based on their chemical structure, enabling sequencing. This technique is advantageous for *hachimoji* DNA sequencing because the signal only depends on the chemical structure of the analyte, with no requirements related to interaction with biological systems present in other techniques.

I begin in chapter 1 with an introduction into artificial genetic systems, their applications, and the specific difficulties involved in sequencing them. In chapter 2 I summarize the historical development of nanopore DNA sequencing, setting the stage for the techniques I developed in this work. In chapter 3 I characterize the readability of *hachimoji* DNA with nanopores. I also analyze the error modes in nanopore sequencing of *hachimoji* DNA that are due to interactions between the non-standard bases and the motor enzyme involved in sequencing. In chapter 4, I build off of the preliminary work in chapter 3 to demonstrate the first direct sequencing of an entirely synthetic DNA alphabet. These results provide insights into further development of artificial genetic systems, such as prioritizing compatibility with existing biomolecules. This work is also relevant to the study of "natural" non-standard bases that occur in biology and have relevance to human health. These include DNA methylation and RNA modifications that have so far proven difficult to interrogate.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AEGIS: Artificially Expanded Genetic Information System. Used to describe the 12-letter DNA system incorporating hydrogen bonding base pairs, or a subset of this alphabet.

ALIEN DNA: ALternative Isoinformational ENgineered (ALIEN) 4-letter DNA alphabet consisting entirely of the artificial bases **P**,**Z**,**B**, and **S**.

ATP: Adenosine triphosphate. A nucleic acid used in biological systems as an energy storage molecule. A common fuel source for many enzymes.

CODON: A sequence of 3 bases that code for an amino acid or otherwise regulate protein translation.

DE BRUIJN SEQUENCE: A maximally compact sequence of that contains every substring of given length possible in a given alphabet.

DE NOVO: Latin for "starting from the beginning." *De novo* sequencing implies sequencing from scratch with no reference sequence.

DNTP: Deoxyribonucleotide triphosphate. These are the building blocks of DNA, they consist of a DNA base bound to a sugar and three phosphate groups.

DSDNA: double-stranded DNA.

HACHIMOJI DNA: Eight-letter DNA alphabet consisting of the artificial bases **P**,**Z**,**B**, and **S** in addtion to the four standard bases A,C,G, and T.

HEL308: DNA helicase enzyme used in our studies to control DNA motion through the nanopore.

HYBRIDIZATION: The process in which two ssDNA strands with complementary sequences bind to form dsDNA.

K-MER: String of "k" consecutive nucleotides in a DNA sequence.

K-MER MODEL: The k-mer model (or k-mer map) is the mapping of each possible k-mer to its corresponding nanopore conductance measurement. A typical k-mer size is 4 nt.

MRNA: Messenger RNA. The RNA product of gene transcription that serves as the template for protein synthesis.

MSPA: *Mycobacterium smegmatis* porin A. A bacterial outer membrane pore with ideal qualities for DNA nanopore sequencing.

NEXT-GENERATION SEQUENCING (NGS): Any of a set of new sequencing technologies that have helped to drastically reduce the cost of sequencing since 2008.

NT: Abbreviation of nucleotide

ONT: Oxford Nanopore Technologies: a company that produces nanopore sequencing products.

PCR: Polymerase chain reaction. A widely used technique for replication of specific segments of DNA.

PRIMER: A short ssDNA strand that hybridizes to a template strand. The primer is then elongated by a DNA polymerase during DNA replication.

SSDNA: single-stranded DNA.

ACKNOWLEDGMENTS

I would like to thank my advisor Jens Gundlach for accepting me into the lab and for his vision and perspective. My colleagues Jon, Andrew, Henry, Akira, and the rest all provided exceptional mentorship, collaboration, and comradery. Professor Daniel Fologea graceously provided guidance and caring mentorship when I began my research career. Daniel represents that which I strive to be as scientist and person. I wish to express sincere appreciation to Karan for her ceaseless support and companionship through the years. Karan along with the lab's 300 lb copper plates kept me grounded during this process. Thank you to my brother Derek for introducing me to science, and encouraging me to apply myself beyond what I think myself capable of. Thank you to my parents Karen and Steve Thomas for their constant support. And of course thank you to all my friends and family that gave me community and a fulfilling life outside of the lab.

This work was supported by the National Institutes of Health, National Human Genome Research Institute grants U24HG011735 and R01HG005115.

DEDICATION

To Daniel Fologea

Chapter 1
INTRODUCTION

1.1 Nucleic acids

The double helix deoxyribose nucleic acid (DNA) is one of the most well-known molecular structures, with many non-scientists having basic familiarity with the structure and function of the compound. This is likely because DNA uniquely carries the genetic information of all known living organisms[1].

A strand of DNA is composed of a polymer of repeating monomers of deoxyribose sugars and phosphate molecules, forming the backbone of DNA[1] (Figure 1.1). The phosphate groups confer a net negative charge to the DNA strand of a magnitude of one electron charge per monomer. The backbone has an intrinsic directionality, with a $5'$-end possessing a terminal phosphate and a $3'$-end possessing a terminal hydroxyl group. Attached to the sugars of the backbone are nucleobases of which there are canonically four types: adenine (A), cytosine (C), guanine (G), and thymine (T). The sequential ordering or "sequence" of these bases along the DNA strand is what carries the genetic information of life in the DNA molecule.

The nucleobases can be separated into two species: A and G are larger double-ringed purines while T and C are smaller single-ringed pyrimidines. The bases have the ability to hybridize with their complementary base, with A forming 2 hydrogen bonds with T and G forming 3 hydrogen bonds with C. This hydrogen bonding

[1]Unless one considers RNA viruses to be alive

along with the size complementarity between purines and pyrimidines allows two complementary strands of DNA to hybridize, forming an antiparallel double helix structure. Remarkably this hybridization occurs despite Coulomb repulsion between the two anionic strands. Ribonucleic acid (RNA) is structurally quite similar to DNA, with the only differences being a slightly different backbone sugar (ribose vs 2′-deoxyribose) and using the base uracil (U) instead of thymine. It is widely thought RNA emerged before DNA or proteins in prebiotic Earth, as it is able to both store genetic information as well as catalyze its own replication[2].

Overall, the structure of DNA is well described as an aperiodic crystal, in which the nucleobases within the double helix form a one dimensional crystal lattice devoid of periodicity. Strong arguments have been made that any theoretical genetic molecule must possess an aperiodic crystal structure[3], with at least one argument having been made even before the discovery of the structure of DNA[4].

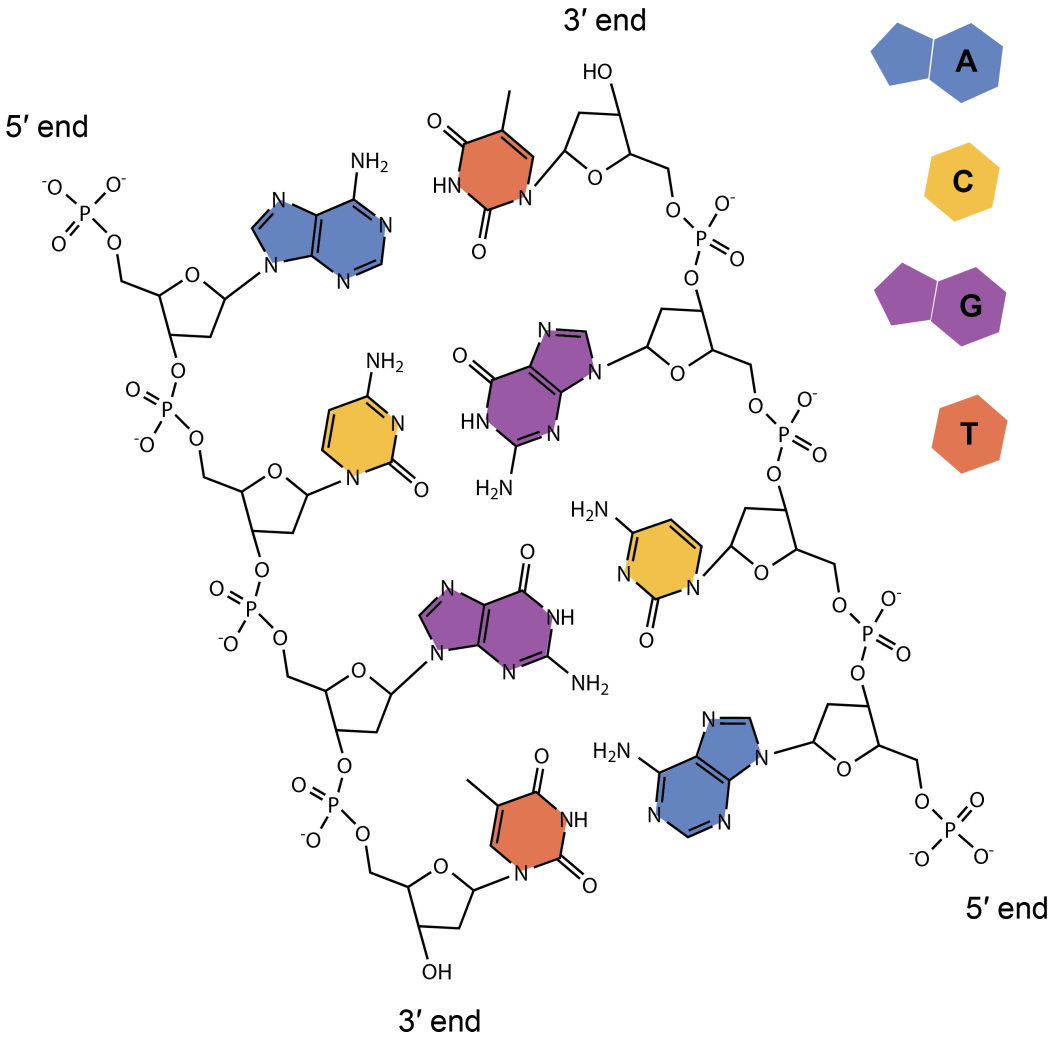

Figure 1.1: Structure of DNA. A single polymer of DNA consists of a sugar-phosphate backbone carrying a uniform negative charge, with each monomer carrying one of four possible nucleobases Adenine, Cytosine, Guanine, or Thymine (A,C,G,T). These form orthogonal base pairs through hydrogen bonding and size complementarity (A:T, C:G), allowing two complementary DNA strands to hybridize into the double helix structure.

1.2 Expanding the genetic alphabet

With at least one interesting exception[5] all known living organisms use the ACGT nucleobase alphabet to encode their genetic information, a universal commonality that points to a single ancient common ancestor. From a chemical perspective, there seems to be nothing unique about these bases. Abiotic sources of purines and pyrimidines found in meteorites contain similar abundances of non-ACGT bases such as xanthine, isoguanine, and purine[6]. Furthermore, modern chemistry has created bases not found in nature but nonetheless still possess the critical properties as the ACGT bases[7]. This all has led synthetic biologists to explore the creation of artificial genetic information systems using non-biological DNA bases (Figure 1.2).

Building off of initial work showing that the hydrogen bond complementarity of canonical bases can be replaced with hydrophobic pairing[8], the dNaM-d5SICS unnatural base pair[9] was developed. The dNaM-d5SICS pair can function alongside the canonical ACGT system, allowing for the creation of a 6-letter DNA alphabet. Independently, another group created the 6-letter artificial genetic system using the 7-(2-thienyl)-imidazo[4,5-b]pyridine (Ds) and pyrrole-2-carbaldehyde (Pa) hydrophobic base pair[10], capable of replication and transcription[11]. However because they lack hydrogen bond pairing, hydrophobic base pairs rely solely on base stacking within the interior of DNA and thus consecutive hydrophobic bases in a sequence may destabilize dsDNA[12].

Another group took a different approach to incorporating non-standard bases into nucleic acids by developing synthetic bases that pair through hydrogen bond complementarity and size complementarity as canonical bases do. In doing so, they developed the Artificially Expanded Genetic Information System (AEGIS) bases[7, 13], which are more structurally similar to the canonical bases than the hydrophobic base pairs.

AEGIS bases leverage hydrogen bonding patterns that are possible yet are "unused" in canonical DNA. With up to three bonding sites in each base (3 in C,G,T,U, 2 in A) of which each site may be either a proton donor or proton acceptor, there are eight possible bonding patterns that a base could adopt. In addition to bond complementarity, bases also use size complementarity (purines bind with pyrimidines) enforced by the size constraints of the double helix. With eight possible bonding patterns and two possible size types (purine or pyrimidine), we now have up to sixteen possible bases (of which ACGT are included) to use in a genetic alphabet. These synthetic bases also use identical pairing rules to the canonical bases, meaning that consecutively incorporated non-standard bases might not disrupt the overall DNA structure. Due to chemical stability constraints, only eight non-canonical bases could be synthesized using this strategy, bringing the total number of AEGIS bases to twelve[14].

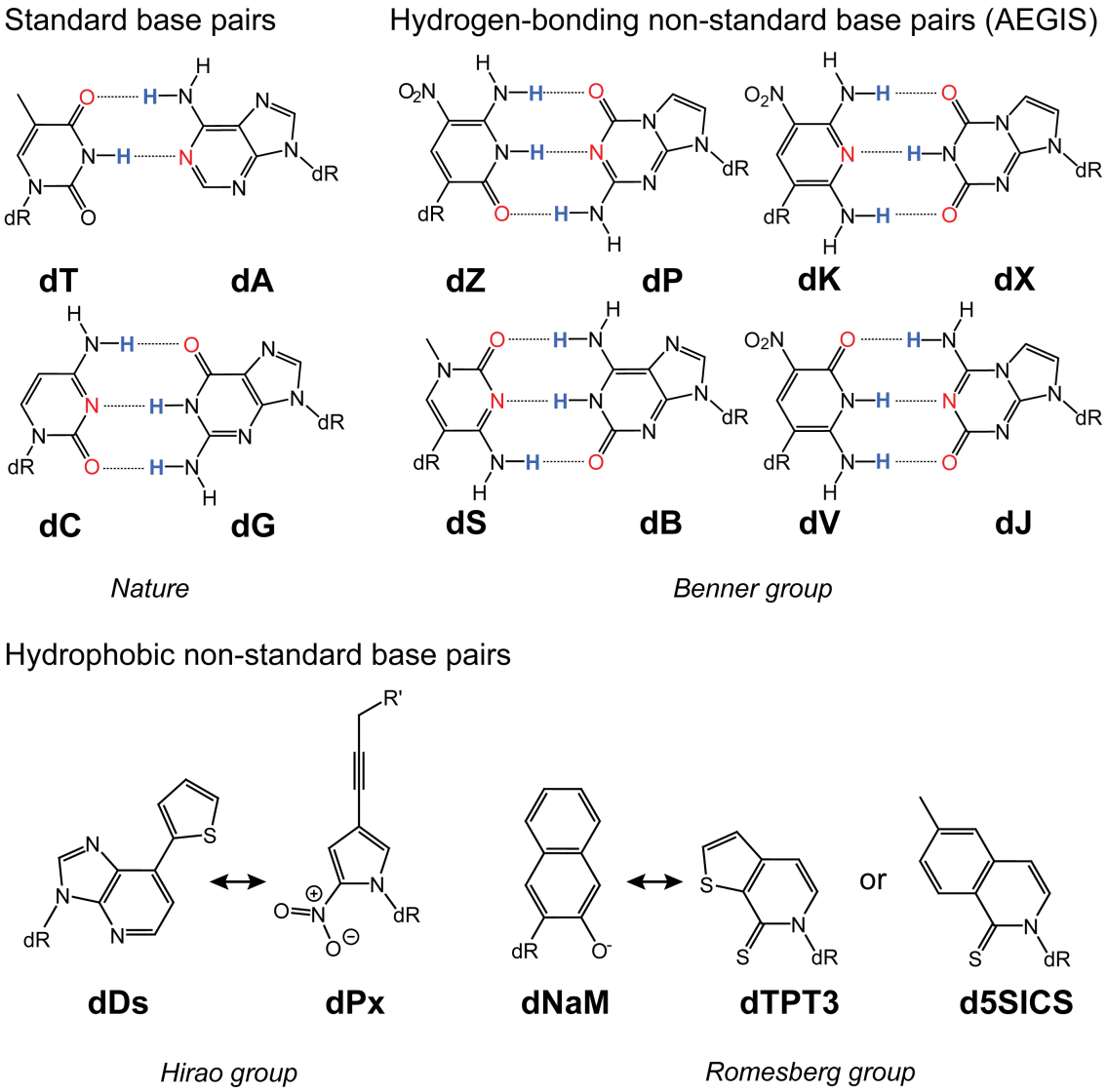

Figure 1.2: Chemical structures of some notable non-standard bases. The standard nucleobases pair through hydrogen bonding and size complementarity, as do the AEGIS non-standard base pairs[13]. Various bases pairing through hydrophobic interactions and size complementarity have also been developed[10, 9]. Dashed lines indicate hydrogen bonding, with the bond donor and acceptor labeled blue and red respectively. R' indicates a modular functional group.

1.3 Applications

Artificial genetic systems retaining the standard DNA backbone but built with non-standard bases have multiple useful properties. One property is that these systems can be supernumerary to the standard alphabet, resulting in a information density greater than canonical DNA. The combinatorics of alphabet size means that while a four base alphabet can produce $4^3 = 64$ unique codons[2], an eight letter alphabet can produce $8^3 = 512$ unique codons. Supernumerary alphabets thus have strong benefits for applications such as molecular barcoding, in which single molecules in a cell or other sample are tagged with unique DNA[3] sequences[15]. Secondly, even short sequences built with non-standard bases can be guaranteed not to hybridize with any section of any genome of any known organism. This feature has utility in amplification of analyte nucleic acids, discussed in the context of viral diagnostics below. Finally, many non-standard bases can prevent the strands they are embedded in from triggering an immune response when within an organism. This has already been used with great effect in the Pfizer-BioNTech and Moderna mRNA vaccines for COVID-19, in which the non-standard base N1-methylpseudouridine was incorporated to shield the mRNA from the immune system while it enters and is subsequently translated by cells[16].

The most mature application of artificial genetic systems is currently in diagnostics, with multiple commercial products on the market. The products include detection of genetic material from viruses such as SARS-CoV-2[17], Zika[18], Dengue, West Nile, and HIV[19]. The advantage that artificial genetic systems provide in these products is the orthogonality of the non-standard base pairs to the

[2]A codon is a sequence of 3 bases that code for an amino acid or otherwise regulate protein translation

[3]or other informatic polymers

natural genetic material tested for. This is manifested in the optimization of polymerase chain reaction (PCR), in which short oligonucleotide primers hybridize with the target genetic material to prompt exponential amplification of the target provided a small threshold of target originally existed in the sample. Problems arise with the incorporation of multiple primers that aim to hybridize with multiple targets, which would otherwise result in more expansive and efficient diagnostic tests. The more primers are included, the more risk of primer-primer interactions resulting in off-target amplification. Designing primers with sections of non-standard bases allows for more specific hybridization[20, 19], as these non-standard bases by definition do not occur in the genome of any organism or virus.

Systematic evolution of ligands by exponential enrichment (SELEX), developed in 1990, promised a new future of tightly binding, highly scalable ligands made of single-stranded oligonucleotides through harnessing evolution in the lab[21]. These oligonucleotide ligands are often referred to as aptamers. The optimism following the development of SELEX was soon tempered by the structural limitations of standard DNA and RNA, due to the constrained four-letter alphabet of these systems lacking neither the combinatorial diversity nor the chemical functional groups of the 20-letter amino acid alphabet that builds proteins. The comparatively recent development of artificial genetic systems containing 6, 8, 10, or 12 possible letters and even limited forms of functional groups has the potential to revitalize the field of aptamer development through SELEX[22, 23, 24] (Figure 1.3), demonstrated with the evolution of a ribonuclease composed of a 6-letter DNA alphabet[25]. Critical to the function of the ribonuclease was the presence of a nitro functional group on the non-standard base **Z**. Further developments in this field may yield genetic alphabets of high combinatorial complexity and a diversity of functional groups to rival proteins, while being much cheaper and more scalable.

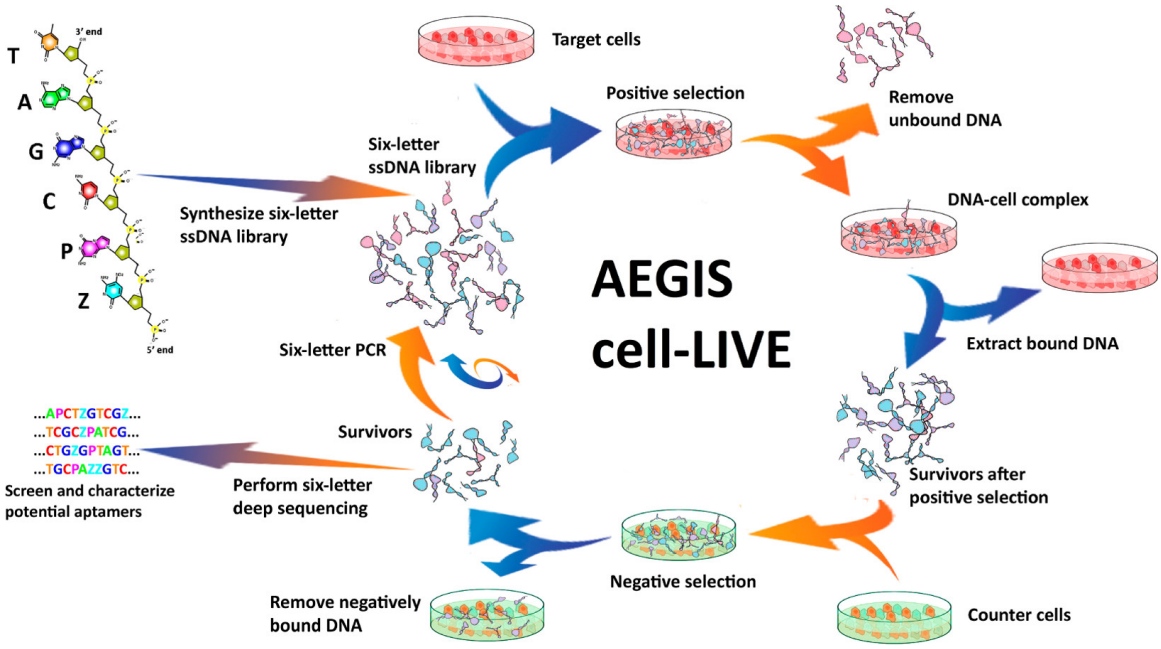

Figure 1.3: Laboratory *in vitro* evolution using AEGIS bases (AEGIS cell-LIVE) can create medically-relevant DNA molecules. In this workflow, six-letter ACGT**PZ** ssDNA was synthesized with a random 25 nt region flanked by primer binding regions. This library was incubated with HepG2 liver cancer cells and any unbound molecules were washed away (positive selection). Molecules which survived were then incubated with healthy liver cells and unbound molecules were kept (negative selection). Survivors were PCR amplified and used for a subsequent round of selection, and a subset were sequenced using transliteration sequencing. This figure was adapted from [26]

By far the most ambitious application of artificial genetic systems is to integrate them into living organisms to create semi-synthetic life[27]. In the last decade, we saw the successful incorporation of a 6-letter alphabet containing the hydrophobic dNaM-d5SICS base pair into *Escherichia coli*. The dNaM and d5SICS bases were stably retained and propagated by the organisms through multiple rounds of replication[28, 29, 30]. During cellular translation of mRNA into protein, a library of 64 3-base codons are used to convert mRNA sequence into amino acid sequence. The same group expanded the amino acid alphabet through an additional synthetic codon that includes the dNaM and d5SICS bases, introducing a downstream functionality to the expanded alphabet[31].

Other applications of artificial genetic systems include the storage of digital information in nucleic acids[32], greatly improved with the increased information density of supernumerary alphabets. Artificial genetic systems also allow for insights into the origin of life and the prebiotic development of the terran ACGT genetic system[33, 34]. With the exploration of that many "possible" forms a biological information storage molecule could take, questions arise about how and why DNA and RNA evolved into their present form.

1.4 The sequencing problem

The process of determining the sequence of nucleic acids is important for any applications involving them. Early DNA sequencing was done with a slow and tedious technique called Sanger sequencing[35] (Figure 1.4 *left*). This is done by using a DNA polymerase to synthesize many copies of the target strand using deoxynucleotide triphosphates (dNTPs) in bulk. dNTPs are the building blocks of DNA, with each free nucleoside coupled to three phosphate groups able to be incorporated into a growing DNA strand by a polymerase. A fraction of dNTPs are

chemically modified to stall polymerase synthesis, and are tagged with a fluorophore specific to the incorporated base. The synthesized DNA is then run on a gel to separate the strands based on polymer length. Upon visualization of the fluorophores in the gel-separated strands, one may determine the sequence of the original strand.

As Sanger sequencing is cumbersome and expensive, there was a strong demand for new sequence technology. This demand was met with the advent of Next-Generation Sequencing (NGS)[36], which drastically increased throughput by massive parallelization (Figure 1.4 *right*). NGS technology varies, but they all use a similar parallelized sequencing-by-synthesis approach. As in Sanger sequencing, DNA polymerase is used to synthesize a complement strand to the target DNA strand with fluorescently-labeled dNTPs. Instead of terminating the synthesis of the strands, the dNTPs allow for continued synthesis of the strand once a moiety is removed chemically in a flow cycle. The fluorescence signal is then read out by sensitive cameras to determine the target sequence.

Adapting NGS or Sanger sequencing directly to artificial genetic systems has not been attempted to our knowledge. Doing so would require development of new fluorescent dNTPs with fluorophores that emit wavelengths distinct from each other and distinct from the fluorophores used for ACGT. This would be coupled with the need to optimize the sequencing polymerase for each genetic system as well as camera and/or optics modifications to allow discrimination of the expanded fluorophore color chemistry. While theoretically possible, this endeavor is prohibitively expensive for a single lab and there is insufficient demand to prompt a company to invest in its development.

To get around this lack of direct sequencing techniques, the transliteration strategy[38] (Figure 1.5) was developed to predictably convert non-standard bases

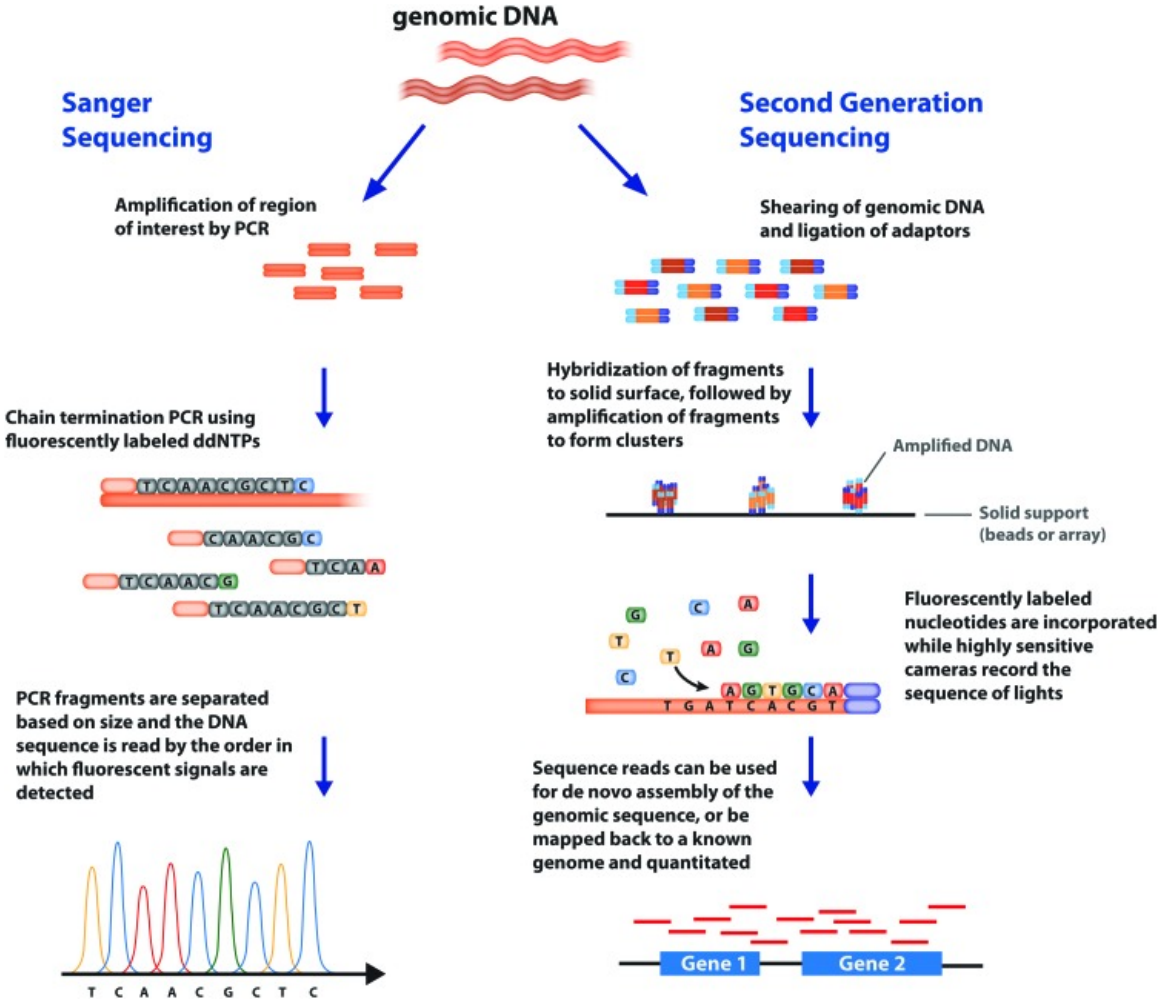

Figure 1.4: Sanger sequencing and next-generation sequencing (NGS). Sanger sequencing (left) begins with the fragmentation and PCR amplification of genomic DNA followed by another PCR step in which elongation is terminated by a fluorescently labeled dNTP. Fragments are then separated by size selection and sequence is determined via the fragment length and fluorescent color. NGS (right) also involves genomic fragmentation, but is followed by attachment to a solid surface. Cameras record as polymerases incorporate fluorescent dNTPs without termination and the fluorescent signal is used to determine sequence. This figure was adapted from [37]

.

into standard bases which are then sequenced using standard Sanger sequencing. This process works by leveraging the protonation state of the non-standard bases by modulation of pH, changing the hydrogen bonding pattern and resulting in predictable mismatches between the non-standard and standard bases. Through this, indirect sequencing of the 6-letter ACGT**PZ** alphabet has been achieved. Caveats of this technique are that a polymerase is required capable of accurately incorporating all non-standard bases[4], pKa's of the non-standard bases must be close to neutral, and the technique is indirect and has low throughput.

[4]Currently ACGT**PZ** is the only AEGIS with a compatible polymerase

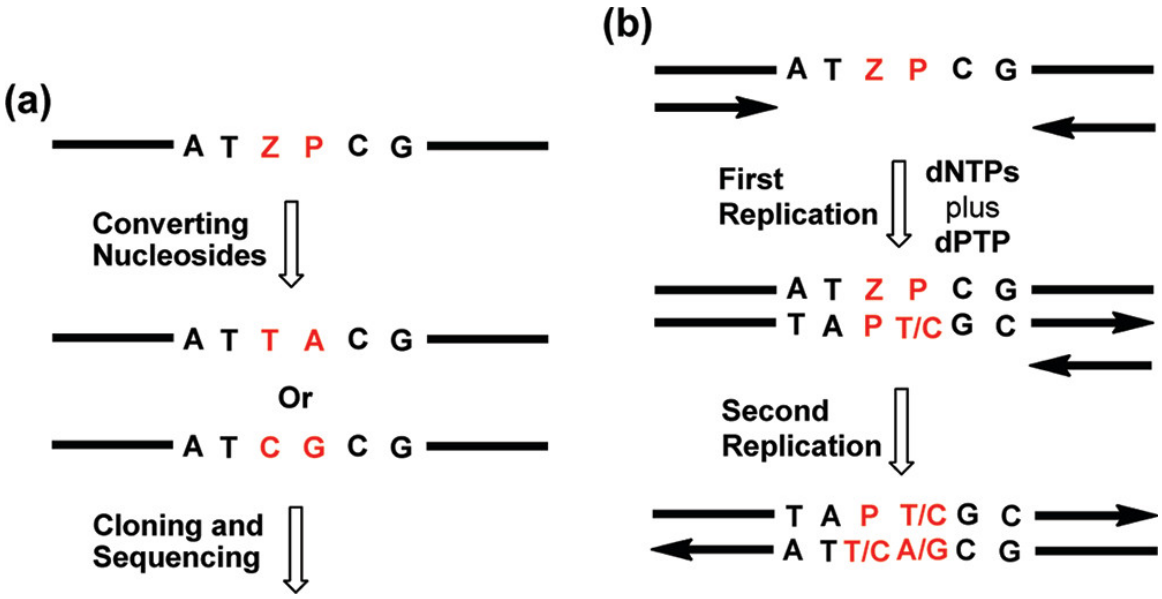

Figure 1.5: Transliteration strategy for sequencing ACGT**PZ**. (a) **Z** and **P** nucleotides within a sequence are converted into a mixture of T:A and C:G pairs using a process based on pH-dependent base pair mismatches. The converted sequences are then sequenced using Sanger sequencing and mixtures of T:A and C:G pairs infer the position of original **Z**:**P** pairs. (b) Manipulating the concentrations of d**P**TP and d**Z**TP enables stepwise conversion of **Z**:**P** pairs into T:A or C:G pairs. This figure is adapted from [38].

1.5 Nanopore sequencing

Nanopore sequencing has gained momentum in the last decade since the first demonstration in our group[39, 40], with commercial devices such as the Oxford Nanopore Technologies (ONT) MinION enabling high-throughput sequencing in real-time at the single molecule level[41]. As it does not require fluorescently-labeled dNTPs or enzymatic synthesis, nanopore sequencing is perhaps the most well-suited of the available sequencing techniques for application to artificial genetic alphabets. This thesis will detail our group's efforts to apply nanopore sequencing to artificial genetic systems.

Chapter 2

NANOPORE SEQUENCING

2.1 Basic concept

Nanopore sequencing starts with a single nanometer-scale pore (nanopore) electrically linking two electrolytic reservoirs (Figure 2.1). An applied electric potential between the reservoirs results in the flow of ions through the pore, constituting an ion current typically at the scale of picoamperes which is measured as the primary signal. Nucleic acids are intrinsically charged molecules due to the negative phosphate groups in the backbone, and thus are electrophoretically pulled through the nanopore by the established electric field. As a strand of NA passes through the pore, it blocks a portion of the ion current. As each type of base has a slightly different chemical structure, they thus block the current to a degree unique to each type of base. This correlation of the type of base to the degree of measured ion current blockage enables the determination of the sequence of the translocating strand.

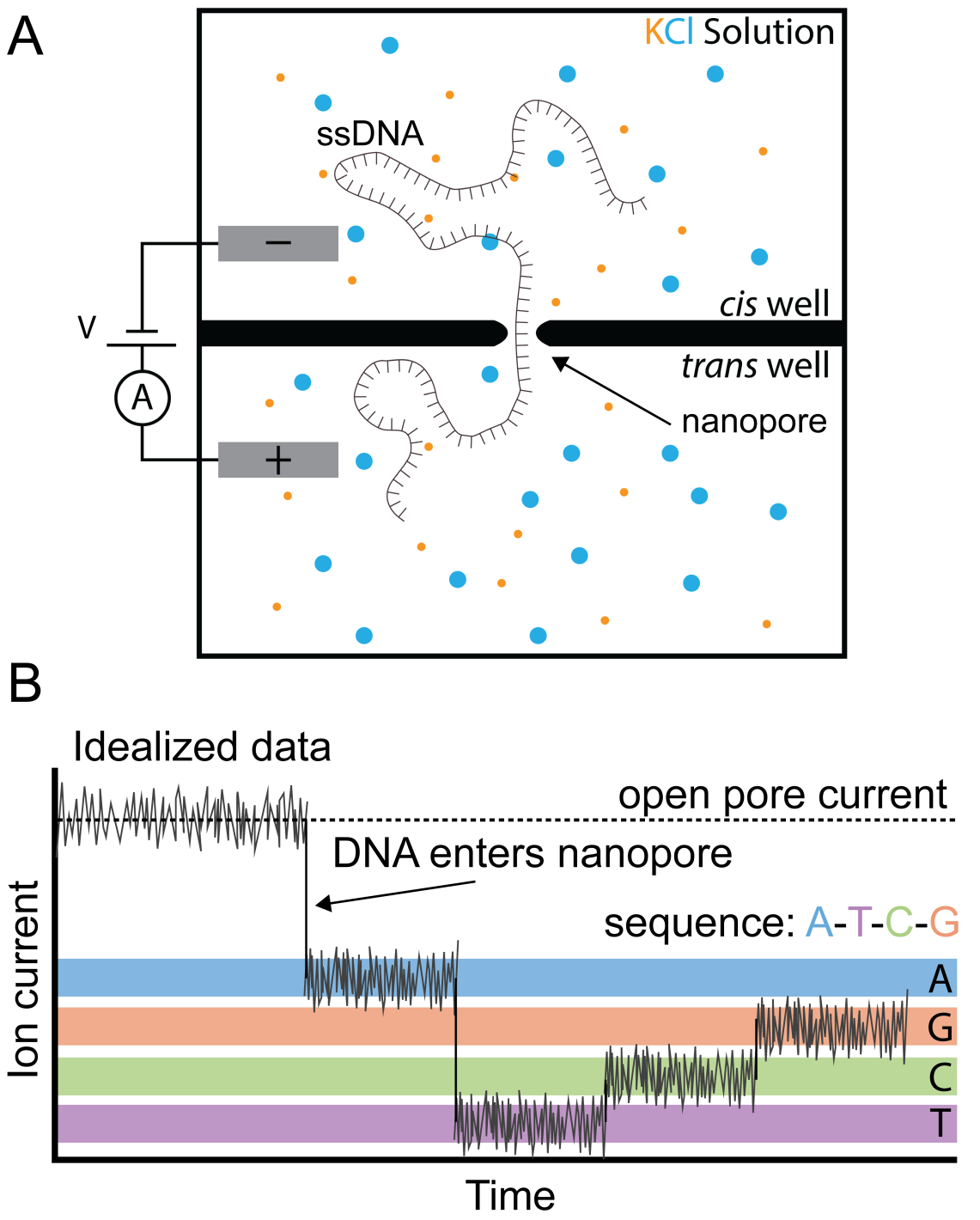

Figure 2.1: Nanopore sequencing basic scheme. (A) A voltage is applied across a single nanoscale pore (nanopore) embedded in a dielectric membrane, establishing an ion current via the electrophoretic movement of ions (K^+ and Cl^-) through the pore. Negatively charged DNA is threaded into the pore by the voltage and blocks the measured ion current to a degree based on the nucleobases in the pore. (B) Idealized ion current trace showing an open pore current and subsequent DNA capture, along with base-specific current blockages that allow identification of the DNA sequence.

2.2 Choosing a nanopore

Nanopores can be broadly separated into two main classes: solid-state and biological. Solid-state pores are typically manufactured through etching a hole through a thin dielectric material, commonly silicon nitride or graphene. This class of pores has the advantage of robustness and easy integration into integrated circuits, but lack structural consistency[42].

As the other main class of nanopores, biological pores are protein pores derived from natural organisms[1]. These pores, unlike solid-state pores, are atomically reproducible, meaning each pore is exactly the same barring misfolding or degradation. Pioneering work with biological nanopores were carried out with α-hemolysin , a pore-forming toxin[44, 45] (Figure 2.2). These pores often must be embedded into a phospholipid bilayer membrane to function, as in their natural context.

Mycobacterium smegmatis porin A (MspA)[46] showed promise as an excellent pore for sensing and sequencing nucleic acids[47]. This is due to its short (0.6 nm) and narrow (1.2 nm) constriction, over which nearly all of the voltage drop between reservoirs occurs. The narrow constriction also happens to be roughly the same width as the cross section of a single-strand of NA, meaning that the NA-specific ion current blockage is large.

Our group has engineered MspA to optimize its sequencing prowess[47]. Wild-type MspA integrated into a nanopore setup has a low NA capture rate, presumably due to many negative charges at the pore rim which repulses the negative-charged NA. By mutating the amino acid residues that carry these charges into residues which hold a positive charge, our group demonstrated a much higher NA capture rate.

---

[1]Recent progress has been made in the *de novo* design of protein nanopores[43]

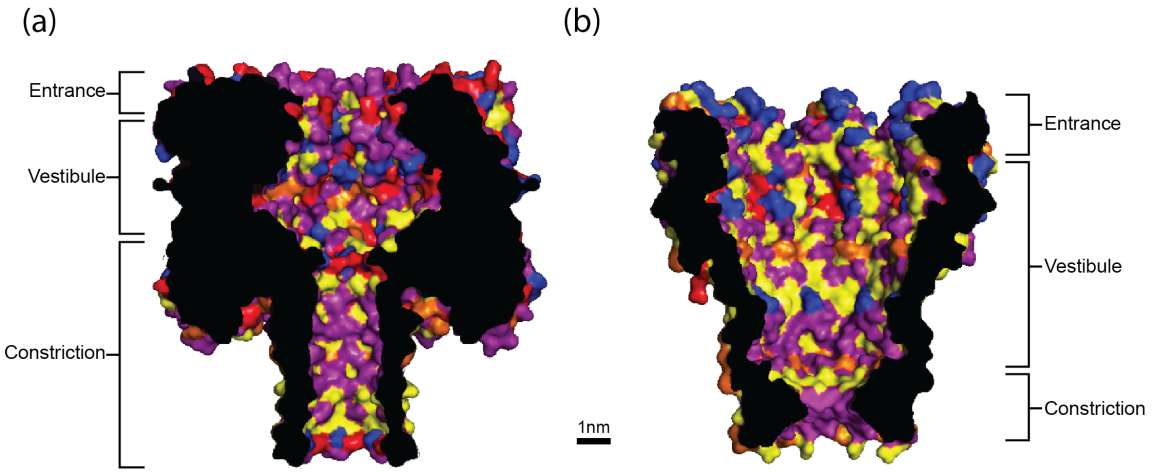

Figure 2.2: Comparison of α-hemolysin and MspA structures. (A) Structure of α-hemolysin shows a long and wide constriction compared to the short and narrow constriction of (B) MspA. As nearly all of the voltage drop occurs over the constriction, a shorter constriction leads to smaller length of the translocating analyte being sensed at any given time (i.e. the shorter the constriction the fewer DNA bases influence the ion current at a time). This figure was modified from [47].

2.3 Controlling NA translocation

NA translocation through a nanopore occurs too quickly (∼2 million bases per second) to resolve ion current blockages into the underlying sequence[47], thus methods of slowing down translocation had to be employed. The most successful method to achieve this has been to employ a NA-processing motor enzyme[39] (Figure 2.3). These enzymes, often polymerases or helicases, bind to and walk along NA strands in discrete stochastic steps using chemical energy[48]. The overall speed of a motor's walk varies widely across enzymes and operating conditions.

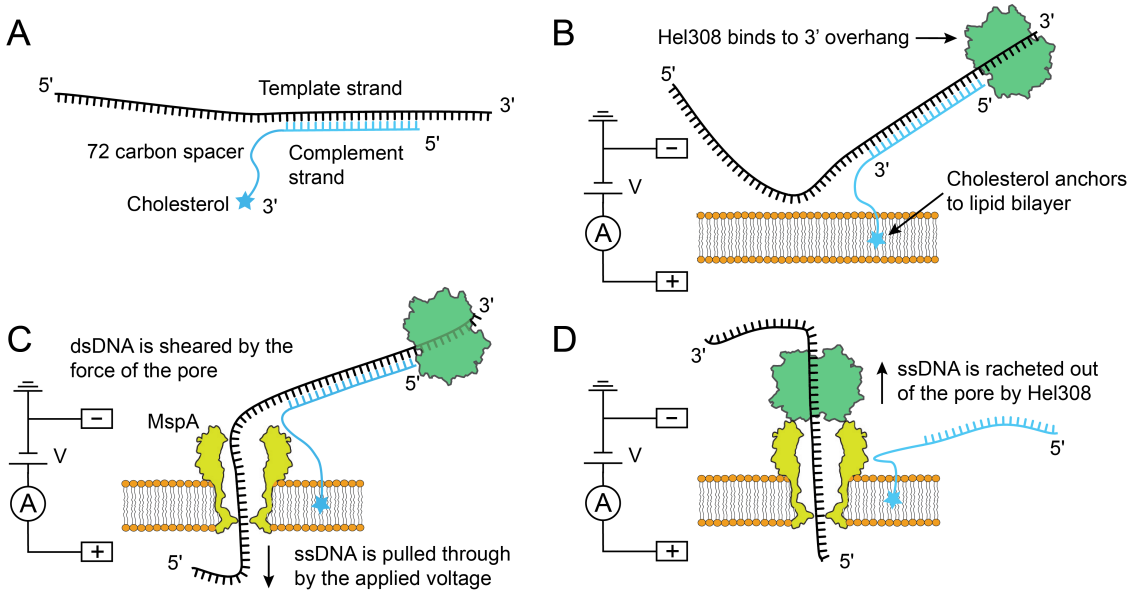

Figure 2.3: Enzymatic control of DNA through the nanopore. (A) Typical synthetic DNA design used for nanopore sequencing consisting of a template strand to be sequenced along with a short complement strand that promotes Hel308 helicase binding at the dsDNA/ssDNA junction (B). The cholesterol at a terminus of the complement strand anchors the complex to the lipid bilayer boosting the local concentration of DNA near the pore. (C) The free 5′ end of the template strand is captured by the electric field near the nanopore and the dsDNA is sheared by the force of the pore. (D) Hel308 comes to rest on the pore rim holding the template strand in place, and pulls the strand back out of the pore in half-nucleotide discrete steps using the hydrolysis of ATP. This controlled stepping of the DNA through the pore enables nanopore sequencing. This figure was adapted from [49].

2.4 Deconvolution of ion current into NA sequence

The signal measured during enzyme-actuated DNA translocation is data rich, and much more complex than four bases would suggest. To determine the resolution of the MspA nanopore system, our group measured a DNA strand containing repeated instances of the trimer 5′...CAT...3′ with one instance swapping the T for a G[39] (Figure 2.4). This sequence produced a repetitive signal except for near the G substitution, in which ∼4 current levels were modified compared to the background signal. Because this experiment was done with phi29 DNA polymerase which walks along DNA in single nucleotide steps, these data show that MspA is sensitive to ∼4 nucleotides simultaneously near the pore constriction.

This suggested that de novo nanopore sequencing could be performed using a 4 nucleotide k-mer model (a map of measurements of all possible permutations of "k" consecutive bases). Our group demonstrated this by constructing a $4^4 = 256$ nt de Bruijn sequence containing every 4 nt kmer (quadromer) and measuring it with MspA and the phi29 polymerase[40]. The resulting kmer model was highly predictive of measured ion current, as seen during measurement and comparison of the phi X 174 genome (Figure 2.5B).

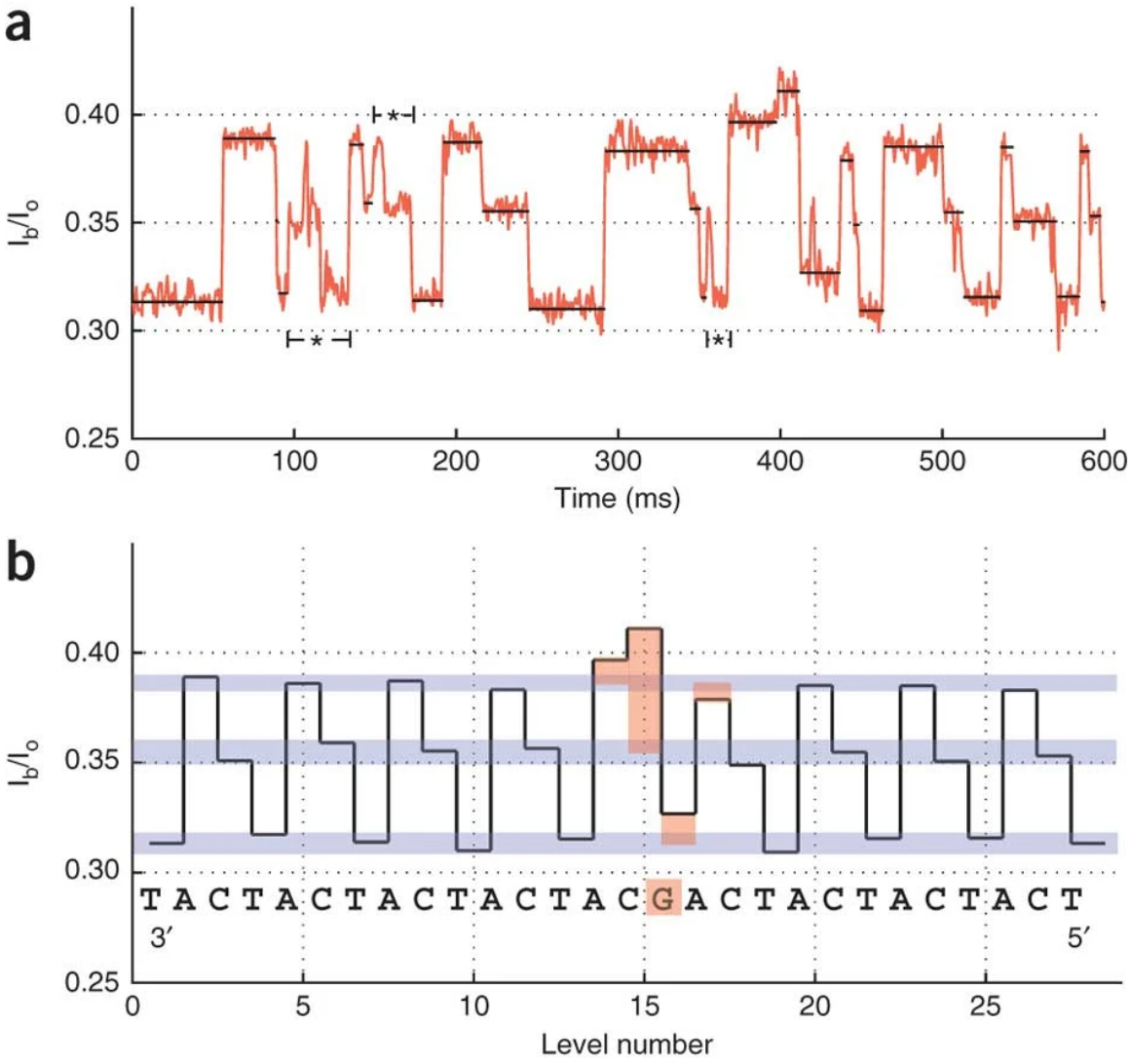

Figure 2.4: MspA sensitivity. (A) Nanopore signal of phi29 DNA polymerase translocating DNA with a repetitive sequence motif of 5′...CAT...3′. The motif is modified to 5′-CAG-3′ in one instance and produces a clear signal difference near 400 ms. (B) The levels of the signal in (A) omitting time information shows the deviation of the modified sequence from the background in orange. The deviation is largest nearest to the base substitution. This figure was adapted from [39].

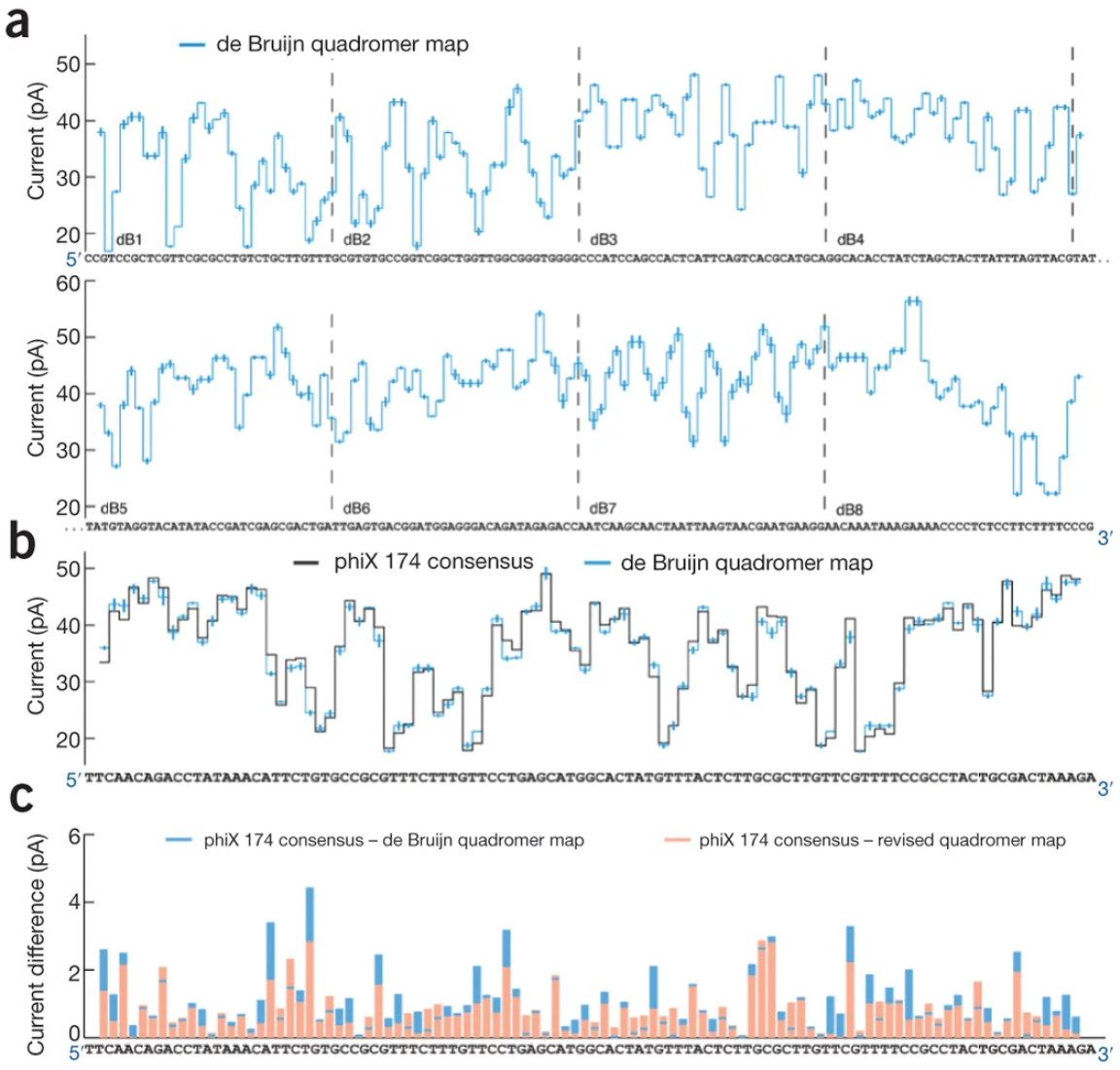

Figure 2.5: First MspA k-mer model of DNA. (A) Consensus currents measured from a de Bruijn sequence containing all possible 4-nt k-mers (quadromers) constitute the k-mer map used for sequencing. (B) The k-mer map prediction of a portion of the phiX 174 genome overlaid with consensus measurements of the genome shows that the map has predictive power. (C) Current difference map of the phiX 174 consensus with two instances of the k-mer map show improvements when additional measurements are added to the map. This figure was adapted from [40].

2.5 Commercial nanopore sequencing

The technology developed by our group held promise for future commercialized nanopore sequencing devices. Based on this, Oxford Nanopore Technologies (ONT) developed the MinION: a handheld device capable of massively parallelized long read nanopore sequencingl[41]. In the past decade, ONT devices have become a mainstream instrument for sequencing and can be found in countless labs around the world. To this day still, the single-read sequencing accuracy of ONT devices cannot approach other NGS techniques such as Illumina sequencers[50], limiting their utility especially in clinical applications.

2.6 Variable-voltage sequencing

In the following years, our group continued to improve nanopore sequencing. First, we replaced the phi29 polymerase with the Hel308 helicase as our motor enzyme. We found that Hel308 steps along DNA in half nucleotide steps[51], allowing for better sequencing resolution compared to the full nucleotide steps of phi29. However the largest innovation came with our replacement of the constant applied voltage with a variable periodic applied voltage during sequencing[52] (Figure 2.6). Because DNA is stretched by the electrostatic force from the applied voltage, applying a time-varying voltage allows the sampling of multiple DNA positions inside a single enzyme step. By applying a 200 Hz 100 mV peak-to-peak triangle wave, we "floss" the DNA back and forth in the pore much faster than the enzyme steps ($\sim$10 Hz). This method extracts more information from the underlying sequence compared with constant voltage sequencing, and transforms the downstream data representation from current levels specified by a mean and error, into conductance states specified by an offset, slope, curvature, and a covariance matrix. The resulting sequencing accuracies are much higher than with constant voltage nanopore sequencing.

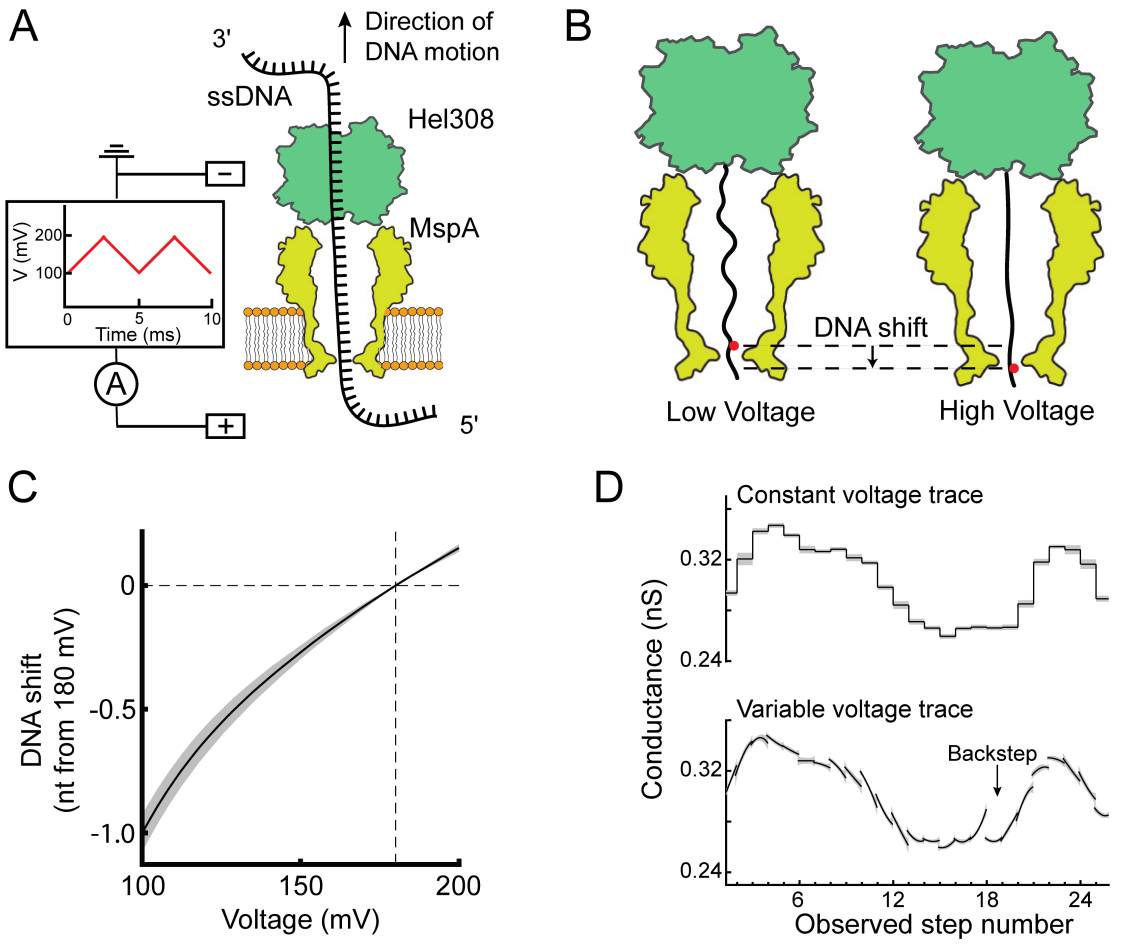

Figure 2.6: Variable-voltage nanopore sequencing. Variable-voltage nanopore sequencing. (A) The experimental setup of variable voltage sequencing is identical to that of constant voltage sequencing (Figure 2.3), except for the application of a 200 Hz, 100 mV peak-to-peak triangle voltage waveform. (B) Different voltages exert different forces on the DNA, causing a shift of DNA position in the pore due to stretching. (C) Measurements of DNA shift as a function of applied voltage relative to DNA position at 180 mV because of the voltage stretching of the DNA section in the nanopore. The shaded area shows the standard deviation of the measurement. (D) Comparison of a constant voltage trace with a variable voltage trace of the same DNA sequence. The variable voltage trace samples more of the DNA strand per observed step, allowing identification of enzyme missteps such as backsteps and ultimately results in higher sequencing accuracies [52].

In constant voltage sequencing, measurement states resulting from discrete enzyme stepping can be well-characterized by the mean current and variance. However variable voltage sequencing states consist of 101-point conductance curves than need additional parameterization beyond mean and variance to accurately characterize. To this end, we use principal component analysis to linearly transform the conductance curves into a basis that represents the largest variation in the data. We found that linear combinations of the first 3 principal component vectors well describe the conductance states while omitting noise (Figure 2.7, Equation 2.1). These principal component vectors roughly map to the mean, slope, and curvature of the conductance curves (Equation 2.2). We would also like to know the variances of these conductance states. Because the measurements are now multivariate, we must estimate the covariance matrix for the 3 variables of a conductance state (Equation 2.3). Because the principal component vectors form an orthonormal basis, this covariance matrix must be diagonal as the variables should have no covariance with the others. However, a non-linear normalization transformation is applied to the 101-dimensional conductance curves (Appendix A), meaning that the principal component measurement variables do covary and the covariance matrices of the final conductance state measurements have off diagonal elements.

$$\boldsymbol{g}_3 = \begin{bmatrix} \boldsymbol{PC}_1 & \boldsymbol{PC}_2 & \boldsymbol{PC}_3 \end{bmatrix}^T * \boldsymbol{g}_{101} \tag{2.1}$$

$$\boldsymbol{g}_3 = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} \tag{2.2}$$

$$\Sigma_g = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \tag{2.3}$$

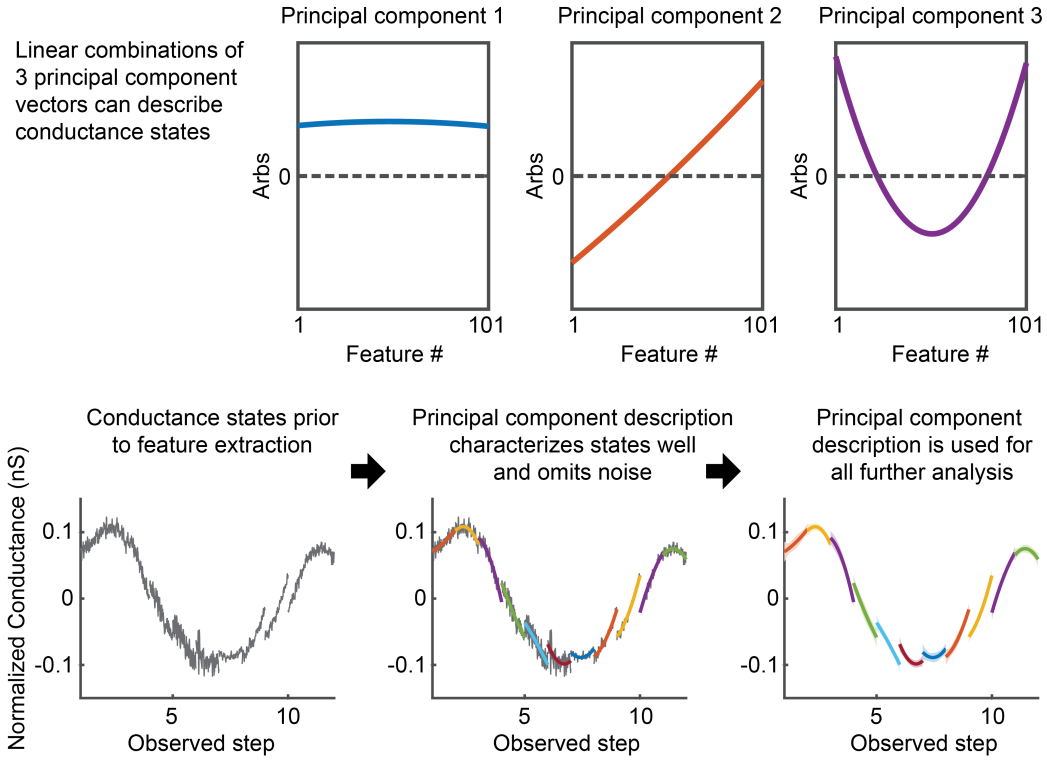

Figure 2.7: Variable-voltage feature extraction. For analysis of variable-voltage traces, we reduce the dimensionality of the conductance vectors of each Hel308 step (state) from 101 elements to 3 using principal component analysis as described in [52]. We thus characterize each state as a linear combination of 3 principal component vectors roughly translating to the offset, slope, and curvature of the signal. Using the principal component description of each state allows us to estimate the covariance of the elements from just 2 voltage cycles, as each half cycle of the voltage waveform can be treated as a distinct measurement. The principal component description of variable-voltage signals is used for all further analysis. This figure was adapted from [49].

Because artificial genetic systems may contain more than four bases[2], the combinatorial explosion of k-mers limits the utility of conventional nanopore sequencing to interrogate these alphabets. Innovations such as variable voltage sequencing that provide more data rich measurements may be key to overcoming this challenge, as constant voltage kmer measurements of large alphabets are guaranteed to have problematic degeneracy due to the limited signal range of the nanopore and the simplicity of the k-mer measurement.

2.7 ONT or MspA?

In the next chapter we will explore the application of variable-voltage MspA nanopore sequencing to the *hachimoji* DNA alphabet. The choice of using a single channel MspA-based nanopore system may seem odd considering the throughput advantages and wide-scale availability of ONT nanopore devices. We consider the higher accuracy and experimental customizability of variable-voltage MspA sequencing worth the trade-offs of lower throughput and lower access. Sequencing supernumerary DNA alphabets may prove much more difficult for ONT constant voltage sequencing, and the chemistry of the ONT devices cannot be changed to accommodate other motor enzymes or buffer conditions. One example of an ONT device used to map pseudouridine bases did so using U-C basecalling errors of standard base algorithms[53]. Such a strategy is likely less effective than lower, signal-level accommodations of non-standard bases into the sequencing model, which is difficult with proprietary ONT software. We will adopt just such a strategy in the next chapter with our low throughput but high accuracy MspA-based sequencing system.

---

[2]Orthogonal hydrogen bonding patterns can support up to 12 bases

2.8 The Hel308 helicase

Our group has characterized the Hel308 helicase from *Thermococcus gammatolerans* using nanopore experiments[51, 54, 55]. We've found that Hel308 has two distinct substates within its ATP hydrolysis cycle, cycling between an [ATP]-dependent and an [ATP]-independent step. These substates map to the physical steps taken by the helicase along DNA, with each physical step spanning a half-nucleotide in length and alternating between the kinetic substates[51]. We have previously found that Hel308 is independent of the magnitude of voltage applied in nanopore experiments between ∼30-65 pN[54], and that it exhibits sequence-dependent kinetics[55]. The half-nucleotide stepping of Hel308 allows for more fine-grained sampling of the DNA in nanopore experiments, as measurements are made twice as often along the strand compared to the Phi29 DNA polymerase. This combined with its stochastic stepping velocity of ∼10 Hz means that Hel308 is well suited for nanopore sequencing and was used for the initial development of variable-voltage sequencing.

2.9 Nanopore tweezers

In enzyme-actuated nanopore sequencing, a motor enzyme controls DNA translocation through the nanopore by stepping along the DNA strand. While this experiment is most obviously performed in order to ascertain the sequence of an unknown DNA strand, it can also be used with a known DNA strand to investigate the behavior of the controlling enzyme. This is because the current levels measured in nanopore sequencing are a high-resolution single-molecule record of DNA position, which can be mapped to the absolute position of the motor enzyme on the strand [51, 56] (Figure 2.8,B.3).

Our group developed this technique, called nanopore tweezers, over the last decade and to date have used it to study many different enzymes including helicases and

reverse transcriptases [51, 54, 55, 56, 57, 58, 59, 60, 61]. Nanopore tweezers can resolve sub-Angstrom motions of nucleic acids on millisecond timescales, outclassing the resolution of similar techniques such as optical tweezers, magnetic tweezers, and single-molecule FRET[56].

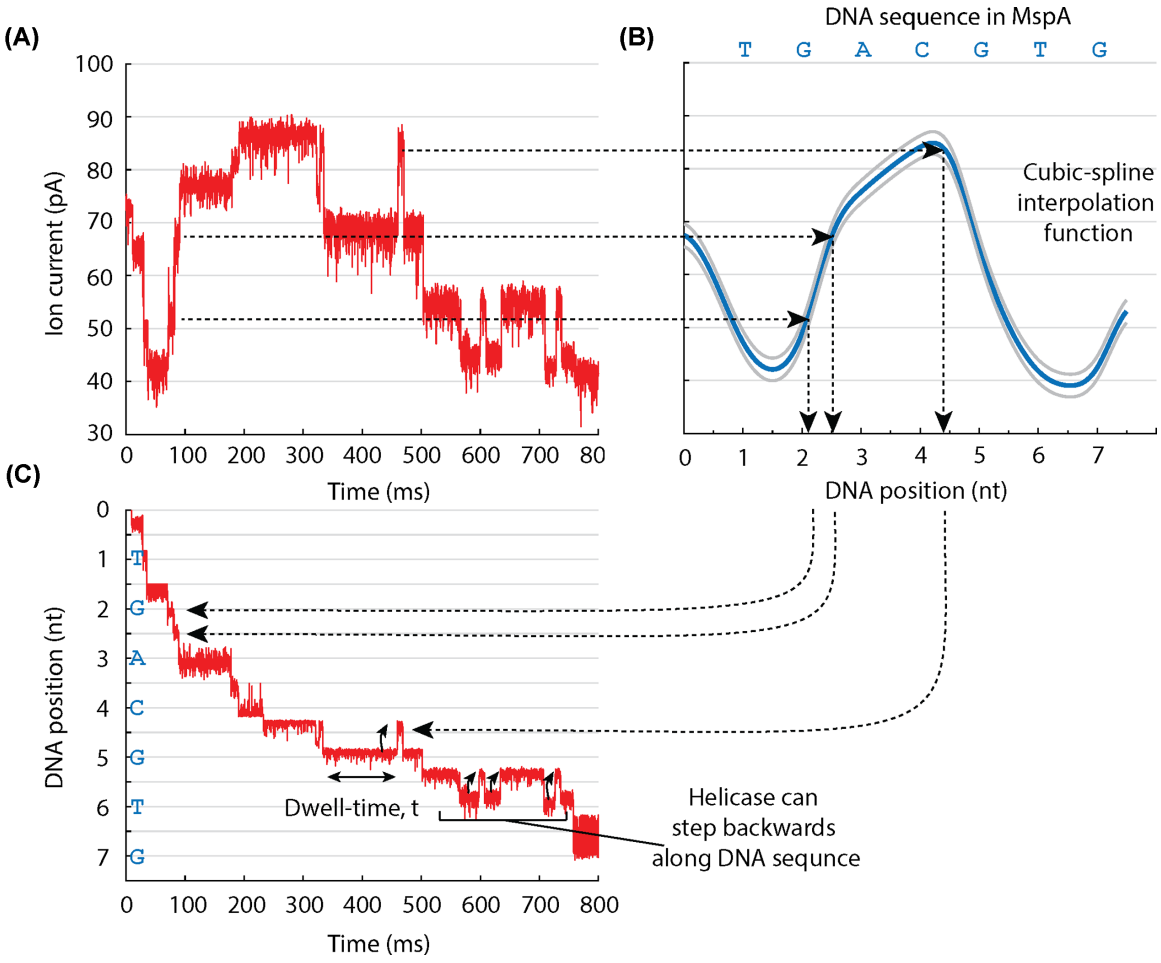

Figure 2.8: Conversion of ion current into DNA position. (A) In nanopore sequencing, the ion current is measured while a motor enzyme samples the NA position in discrete steps, allowing for time averaging of the current for each step. (B) If the enzyme were to continuously feed the NA through the pore, a smooth ion current curve would result. The measured ion current levels can be interpolated using a cubic-spline to estimate this underlying curve. (C) This curve can be used to convert current measurements as a function of time into enzyme position as a function of time. Enzyme kinetic observables can be measured from the data, including dwell-time, backwards steps, and enzyme dissociation from the strand. This figure was adapted from [56].

Chapter 3

ASSESSING READABILITY OF *HACHIMOJI* DNA WITH NANOPORES

3.1 Sequencing hydrophobic base pairs

Our group began its foray into the sequencing of artificial genetic systems through the direct nanopore detection of the hydrophobic base pairs dNaM and d5SICS[62]. Craig *et al.* found that single-base substitutions with dNaM and d5SICS in DNA produced distinct ion current differences. This work led directly to our group using variable-voltage nanopore sequencing on single-nucleotide polymorphisms (Figure 3.1) to measure the replication efficiency of the dNaM and dTPT3 base pairs in the genome of a semi-synthetic organism[30].

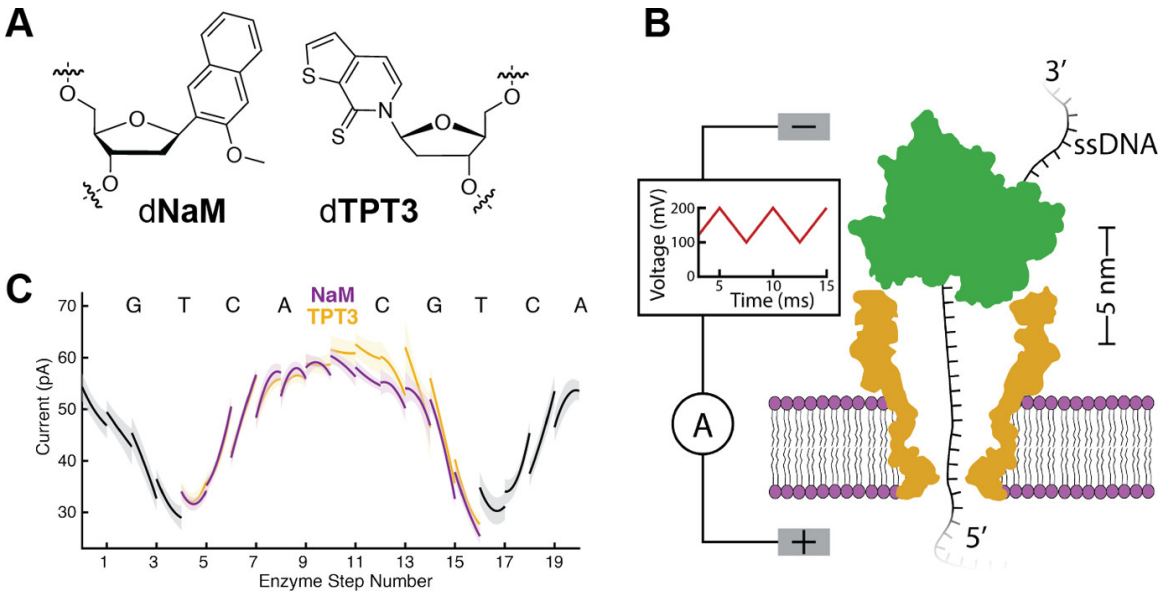

Figure 3.1: Hydrophobic base pair sequencing. (A) The dNaM–dTPT3 UBP. (B) The MspA/Hel308 nanopore system. The MspA porin (gold) is embedded in a lipid bilayer (purple). The Hel308 helicase (green) draws ssDNA through the porin as a variable voltage is applied across the membrane and the current is measured. (C) Consensus current patterns from 40 control measurements of the indicated sequences. Shaded areas represent ±1 SD, demonstrating that single-molecule measurements of dNaM and dTPT3 can be identified with high confidence. Typical base-calling accuracies using variable-voltage sequencing are >90% for the eight hypotheses (the six-letter alphabet plus the resulting abasic site and single nucleotide deletion. This figure has been adapted from [30]

3.2 Hachimoji characterization using nanopores

The work discussed in the remainder of this chapter was published and reproduced with permission from Thomas, CA. *et al.* Assessing Readability of an 8-Letter Expanded Deoxyribosenucleic Acid Alphabet with Nanopores. *Journal of the American Chemical Society,* 145(15):8560-8568, April 2023.[49]

We next turned our attention to an alternative genetic system developed by the Benner group at the Foundation For Applied Molecular Evolution (FFAME). Here, we set out to address challenges that might arise when attempting to sequence *hachimoji* DNA with nanopores. Since *hachimoji* bases have chemical structures similar to standard bases, it was not obvious that nanopore signals from **P**, **Z**, **S**, and **B** would be sufficiently distinguishable from standard base signals to allow sequencing.

Further, decoding sequences requires current values associated with various k-mers to be sufficiently distinguishable. The difficulty of nanopore sequencing scales with the number of base letters in a genetic alphabet (e.g. a 6-base k-mer model requires decoding 4096 unique k-mers for a four-letter alphabet and 262,144 k-mers for an eight-letter alphabet). Sequencing of such an alphabet requires empirical measurements of each k-mer in multiple sequence contexts to develop a current-to-sequence model, requiring significant laboratory time and resources for library construction and analysis. In addition, the ion current levels of these k-mers will have significant overlap with one another within the limited dynamic range of the nanopore ion current, possibly making unique k-mer discrimination intractable. For these reasons, sequencing of base modifications in DNA (e.g. methylation) and RNA (e.g. pseudouridine) has often relied on comparisons of nanopore signals to similar reads of unmodified NA rather than *de novo* detection of the modified

nucleotide[63, 64, 53]. The "k-mer explosion" created by the expansion of the DNA alphabet necessitates novel nanopore technologies capable of extracting additional signal features beyond average ion current, such as variable voltage sequencing[52, 30]. Additionally, nanopore sequencing of unnatural nucleotides requires an enzyme capable of processing them without introducing sequencing errors.

Prior to the significant investment required in building a nanopore k-mer model for *hachimoji* DNA, we assessed whether the four added nucleotides have a distinguishable effect on the nanopore signal. We must also assess their compatibility with sequencing motor enzymes. The ONT sequencing devices are commercially available and have high data throughput and long reads[65]. Despite these advantages, ONT platforms use only ion current data from a single applied voltage, limiting their utility for sequencing expanded alphabets. The motor enzyme is proprietary, highly optimized for sequencing of canonical bases, and cannot be easily modified/optimized for expanded DNA alphabets. Further, ONT from time to time changes its enzymes and pore to optimize standard-base sequencing, and that change may impact the reading of non-standard nucleotides in an unpredictable way.

For these reasons, we chose to investigate the feasibility of sequencing *hachimoji* DNA with variable-voltage sequencing on the MspA nanopore. We measured the nanopore ion current conductance of *hachimoji* DNA and tracked the translocation of the Hel308 motor enzyme (commonly used in nanopore sequencing[66, 30, 62, 52]) to assess enzyme/*hachimoji* compatibility. We also demonstrated proof-of-concept *hachimoji* reference-sequencing by distinguishing single base substitutions of each *hachimoji* base with high confidence using variable-voltage nanopore sequencing.

We first measured the conductance signal of seven *hachimoji* ssDNA templates containing a 16 nt region of homopolymer A, T, C, **P**, **Z**, **B**, or **S** passing through

the MspA nanopore under a constant applied voltage (Figures 3.2, B.1, and B.2). Homopolymer G was omitted because of its tendency to form robust secondary structures, though we did opt to include homopolymer **B** despite its own tendency to form quadruplex and pentaplex secondary structures[67, 68]. These structures may be disrupted by the electrostatic force across the pore during measurement, though they may still contribute to larger noise and wider conductance distribution relative to the other homopolymers. The 3′-nitrogen of the **Z** heterocyclic ring is known to have a pKa of 7.8[25] in free solution, very close to the pH used in these experiments (pH 8). We speculate that protonation of **Z** may be related to the wide conductance distribution measured for homopolymer **Z**, as protonation kinetics are known to be observable as noise in nanopore experiments[69].

Homopolymer **P** gave the lowest measured conductance of the alphabet (0.22 ±0.01 nS), while homopolymer **Z** gave the highest (0.68 ±0.03 nS). The expanded conductance signal range compared to the standard alphabet is encouraging for further sequencing applications, as it suggests *hachimoji* k-mers may be spread out over a larger range of conductance and thus easier to distinguish. The nanopore conductance signal from *hachimoji* homopolymers occupies a larger dynamic range of conductance than the standard DNA alphabet. This increased dynamic range means that additional information can be gleaned from the nanopore signal despite the added difficulties of sequencing 8-letter DNA. In particular, the high currents produced by Z and low currents produced by **P** suggest that *de novo* sequencing of a 6-letter alphabet including A, C, G, T, **P**, and **Z** is a tractable problem.

We next generated a single stranded DNA *hachimoji* reference library with each template having a 65-mer sequence identical across templates except for a single nucleotide which contained one of the eight *hachimoji* bases. We constructed a variable-voltage consensus conductance pattern for each of the templates by

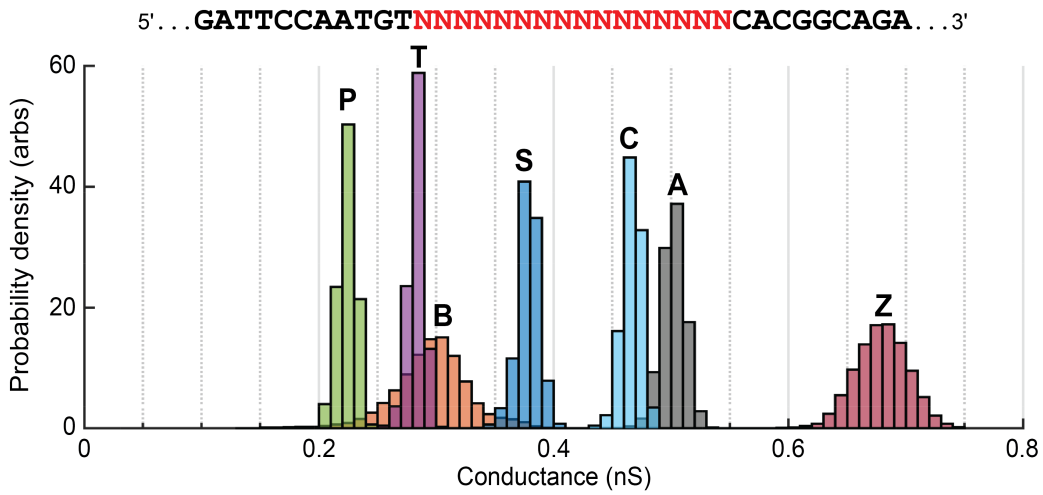

Figure 3.2: *Hachimoji* homopolymer conductance survey. Histograms of constant-voltage conductance caused by translocation of DNA substrates containing 16 nucleotide homopolymers of each *hachimoji* base (excluding G). Homopolymers of **P** and **Z** give the lowest and highest conductance signal respectively, evidence that the *hachimoji* system has an expanded nanopore signal range relative to the standard alphabet.

averaging the patterns of 50 randomly selected variable-voltage reads (Figure 3.4). We used a previously-developed base-calling algorithm[30](Figure 3.3) to align each read to all eight consensus patterns, generating a likelihood that the read corresponds to each consensus. A read is thus base-called by determining the consensus alignment that produced the maximum confidence. We excluded reads used to construct the consensus patterns from base-calling, as well as reads with a maximum consensus alignment confidence less than 90% (low confidence reads occupy 6% of our data). We achieve an accuracy >90% for all bases except for **S** (89.8%) (Figure 3.4).

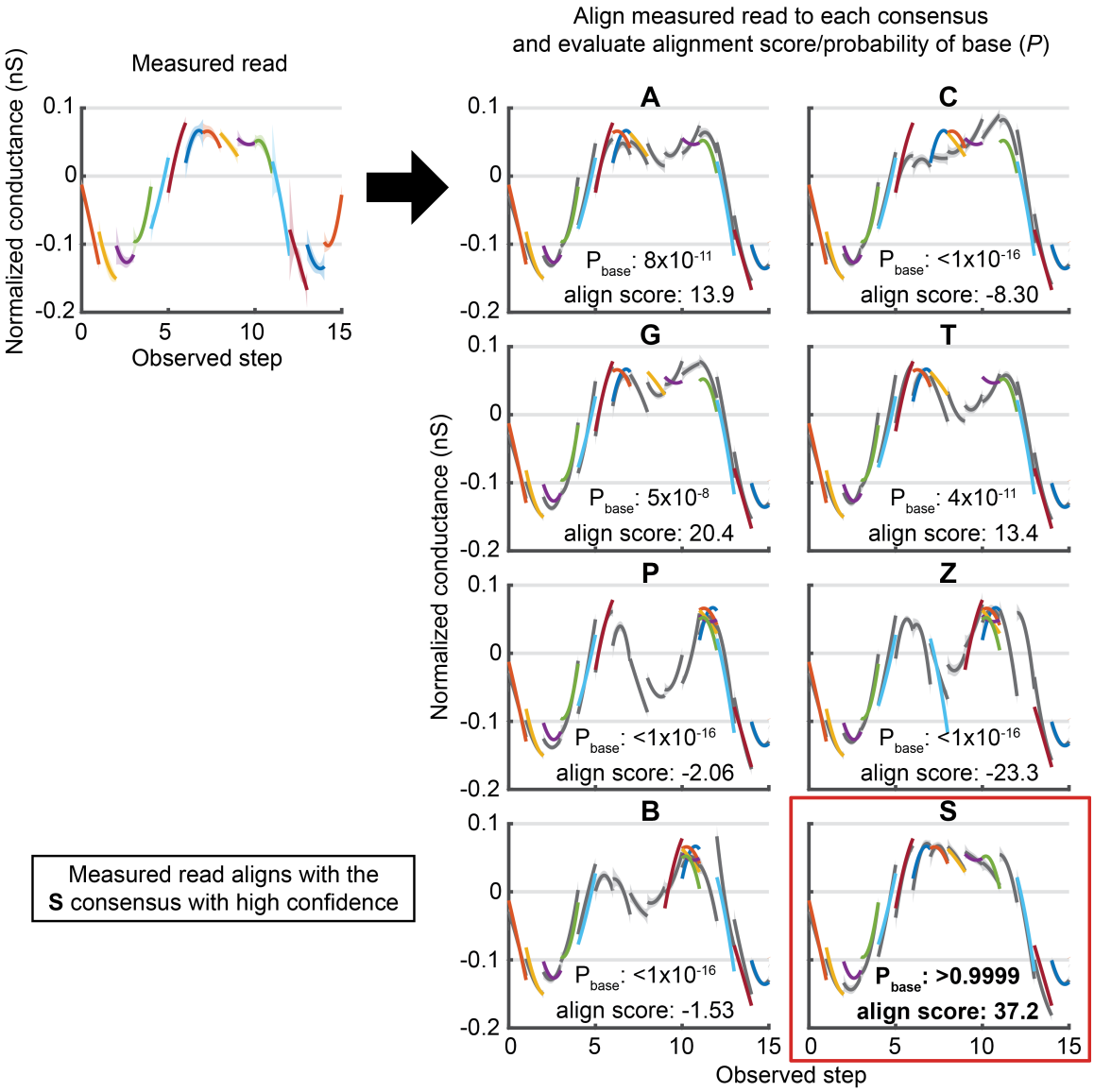

Figure 3.3: Single-base substitution base-calling algorithm. A measured read (left) is aligned to the consensus trace for each of the eight hypotheses (right). The quality of each alignment is scored and is translated into the probability of the read being generated by each hypothesis substitution. This read aligned to the **S** consensus with a probability of 0.99999995 relative to the other hypotheses. The alignments are performed using full length reads, but for clarity only the region influenced by the single base substitution is shown.

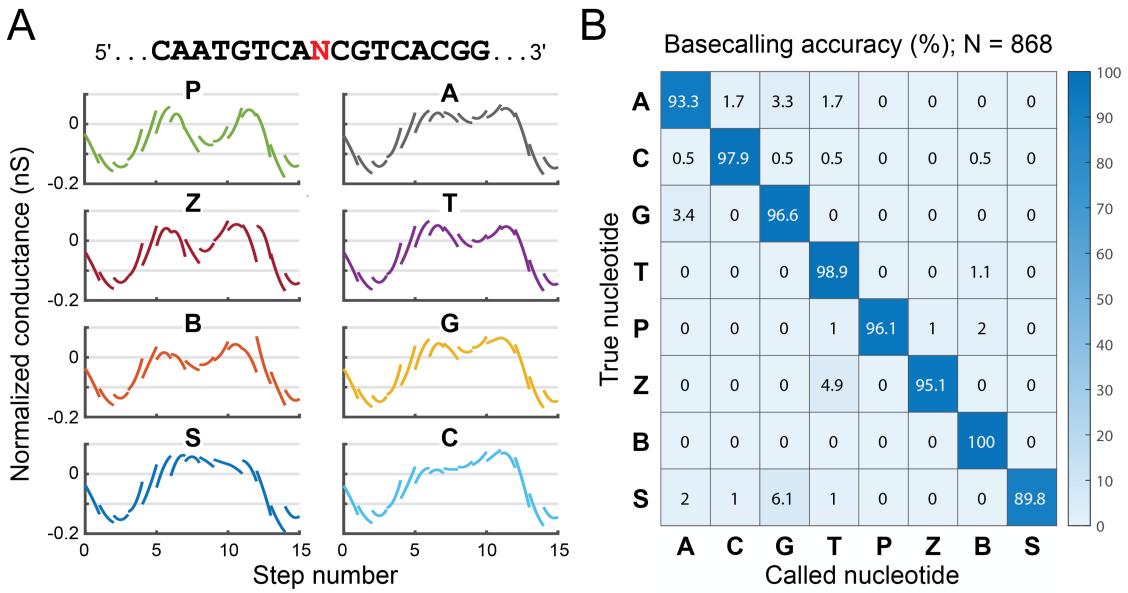

Figure 3.4: Direct reference sequencing single-base substitutions of *hachimoji* DNA. (A) Variable-voltage consensus patterns of *hachimoji* single base substitutions within a pseudorandom sequence. The sequence of conductance-voltage curves from variable-voltage nanopore sequencing contain more sequence information than the sequence of current levels produced from constant voltage sequencing[52, 30]. (B) Confusion matrix of our base-calling algorithm's accuracy shows that *hachimoji* single base substitutions are distinguishable with high confidence using variable-voltage nanopore sequencing. Reads are only included in base calling if they align to one of the consensus patterns with a confidence greater than 90%.

3.3 Hel308 dissociation on homopolymers of non-standard bases

While measuring the conductance of *hachimoji* homopolymers, we observed that many of the reads terminated mid-sequence after processing only ∼10 nucleotides. Considering the measured processivity (average number of nucleotides translocated in a single binding event) of Hel308 is about 1000 nt on genomic DNA[52], this was interpreted as evidence of premature Hel308 dissociation from the DNA strand. As each nanopore sequencing read is a single-molecule record of a motor enzyme's activity, we tracked Hel308 molecules as they translocated on the *hachimoji* templates and measured the position at which Hel308 dissociates from the strand (Figure B.4). We characterized Hel308 dissociation using the Kaplan-Meier estimation[70] of the survival function (Equation B.3). We observed that Hel308 dissociated more often on homopolymers of **S** and **Z** compared to the rest of the templates (Figures 3.5, B.5). **S** and **Z** are connected to the deoxyribose sugar by a carbon-carbon bond (C-glycoside), while **P**, **B**, A, T, C, and G are connected to the deoxyribose sugar by a carbon-nitrogen bond (N-glycoside).

3.4 C-glycosides hypothesis

To test the hypothesis that the C-glycoside character of **S** and **Z** influences Hel308 dissociation, we measured Hel308 dissociation on DNA templates of identical design as described above, but with homopolymer regions consisting of the nucleotides pseudothymidine (C-glycoside analogue of T, "pseudoT"), pseudoisocytidine (C-glycoside analogue of C, "pseudoisoC"), and isocytidine (N-glycoside analogue of **S**, "isoC") (Figure 3.6). Again, C-glycosides resulted in higher Hel308 dissociation (Figure 3.5), indicating that the C- versus N-glycosidic status of the nucleotide appears to be a key determinant of this feature of the Hel308-DNA strand interaction.

Measurements of Hel308 dissociation may be sensitive to the quality of the DNA templates, as helicase dissociation may be influenced by known solid-phase synthesis error modes including DNA fragments or DNA chemical adducts. To rule out template impurity as the main source of our dissociation results, we contracted a private company to perform liquid chromatography-mass spectrometry (LCMS) analysis on a subset of our DNA templates (Figure B.6). This measurement indicated that 19 ± 9% of our DNA molecules were expected to have one or more impurities that might affect Hel308's processivity (Table B.3). This is well below the rate of Hel308 C-glycoside dissociation in which 50-80% of reads ended prematurely and suggests that the glycosidic composition of bases in our DNA templates is the primary cause of the observed differences in the survival curves.

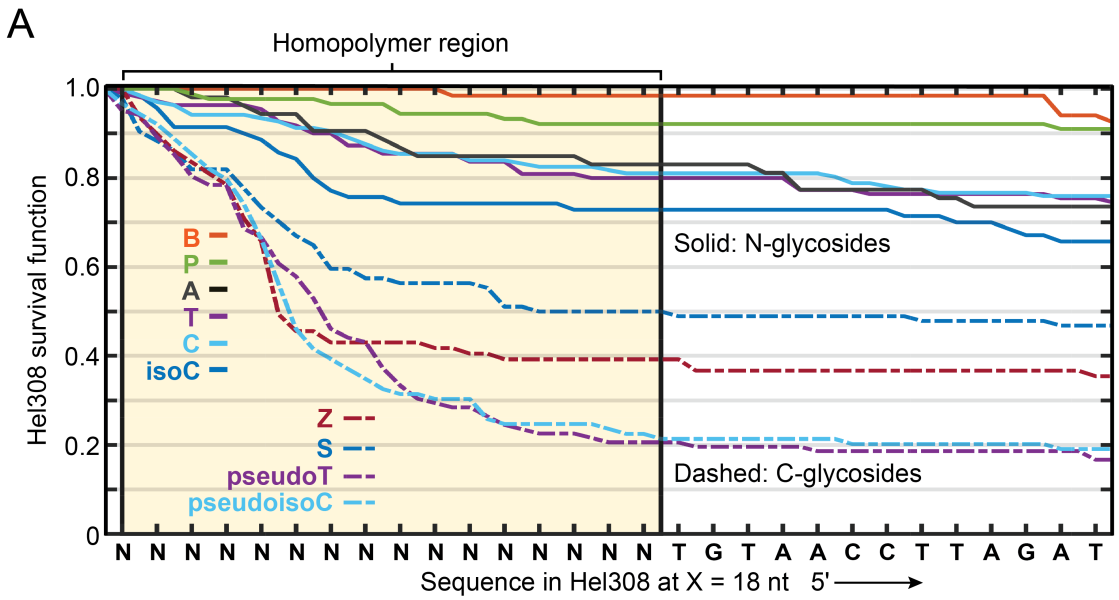

Figure 3.5: C-glycosides prompt early dissociation of Hel308 from DNA. Hel308 survival functions from multiple single-molecule traces as a function of position along a DNA template containing a homopolymer of *hachimoji* bases. C-glycoside nucleosides are associated with higher dissociation (low survival) upon translocation by Hel308. Hel308 steps along the DNA template from 3′ to 5′ in steps approximately half a nucleotide in length.

To further probe the C-glycoside dissociation mechanism, we measured Hel308 dissociation over DNA templates of identical design to those already described but with their homopolymer region replaced with a 16-nt mixed region composed of various combinations of non-standard *hachimoji* nucleotides. Of particular interest is the sequence with a region composed of dinucleotide repeats of each of the four non-standard nucleotides in the pattern: 5′-(**ZZSSPPBB**)x2-3′. The dissociation probability of Hel308 on this sequence is markedly higher over two separate four-nucleotide windows, with the two windows separated by approximately four nucleotides (Figure 3.6). This behavior is consistent with the four-nucleotide-long C-glycoside windows (**ZZSS**) present in the sequence, separated by four nucleotides of N-glycosides. Hel308 dissociation on the sequence 5′-**ZZZZSSSSPPPPBBBB**-3′ also shows high dissociation over a roughly eight-nucleotide window, consistent with the C-glycoside distribution of the sequence. The dissociation probability was related to the sequence at a distance 17 or 18 nt away from the pore constriction in both DNA templates, corresponding to the DNA sequence within Hel308. This distance is similar to distances related to sequence-specific dwell time and backwards steps measured in previous Hel308 studies[55]. As Hel308 is known to contact DNA over a span of ∼14 nucleotides[71], this suggests that the site(s) within Hel308 that most influences C-glycoside dissociation is localized and not distributed throughout the helicase. This also suggests that site-directed mutagenesis may be capable of improving Hel308 survival on C-glycosides.

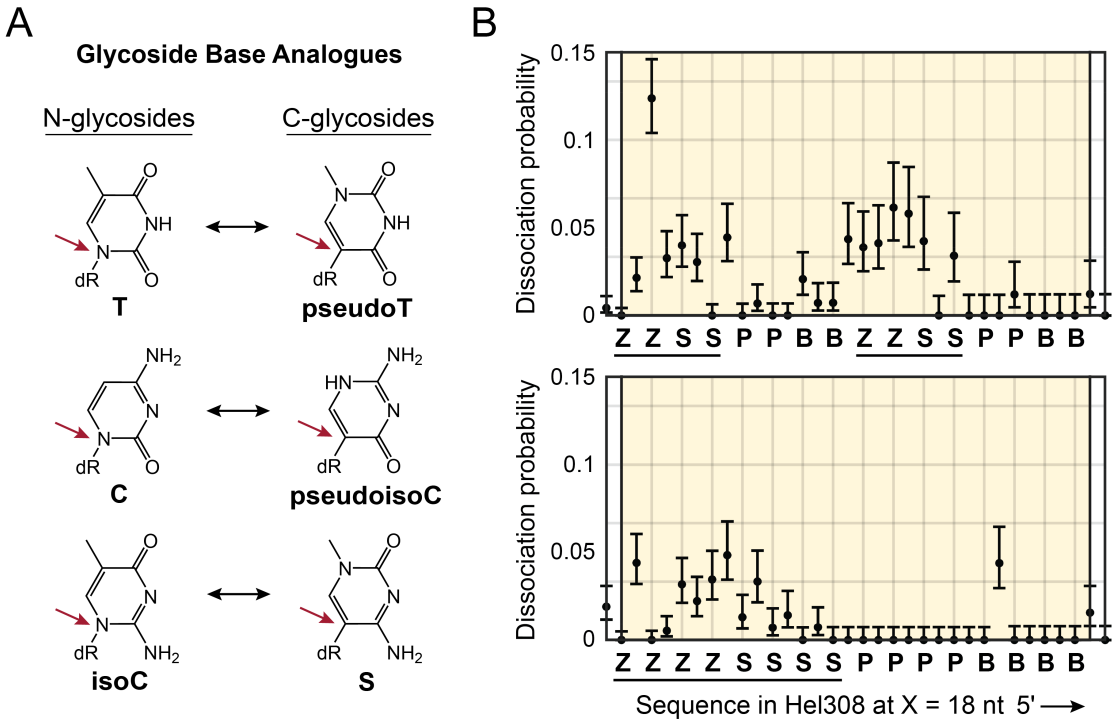

Figure 3.6: Hel308 dissociation on heteropolymers. (A) C-glycosides vs. N-glycosides. Chemical structures of additional nucleotides used to further test the effects of glycosidic character. Bases are connected to the DNA backbone (red arrow) through a C-C bond in C-glycosides while N-glycosides are connected through a N-C bond. (B) Probability of Hel308 dissociation as a function of position on *hachimoji* mixed sequences. High dissociation occurs in windows of approximately four (top) and eight bases (bottom), matching the distribution of C-glycosides (**S** and **Z** underlined) within the sequences. Aligning the dissociation probability to sequence enables coarse-grained localization of the dissociation mechanism to 17 or 18 nucleotides upstream of the pore constriction.

3.5 Non-standard bases elicit sequence specific effects on Hel308 kinetics

In the same experiments characterizing Hel308 dissociation, we also measured Hel308's dwell time and the probability of backwards stepping on the *hachimoji* DNA templates. We have previously measured these kinetic observables of Hel308 on standard DNA and have found they are highly dependent on the underlying sequence[55, 57, 58]. Unsurprisingly, the non-standard *hachimoji* nucleotides also distinctly affect these Hel308 kinetic observables, with **Z** and **B** bases eliciting longer dwell times in most sequence contexts and **P** and **B** causing more backwards steps (Figure 3.7). C-glycosides do not uniquely cause longer dwell times and/or more frequent backsteps in Hel308 compared to N-glycosides, ruling out the alternate hypothesis that increased time spent on the DNA template with a constant dissociation rate is ultimately the cause of the high fraction of dissociations observed on C-glycosides.

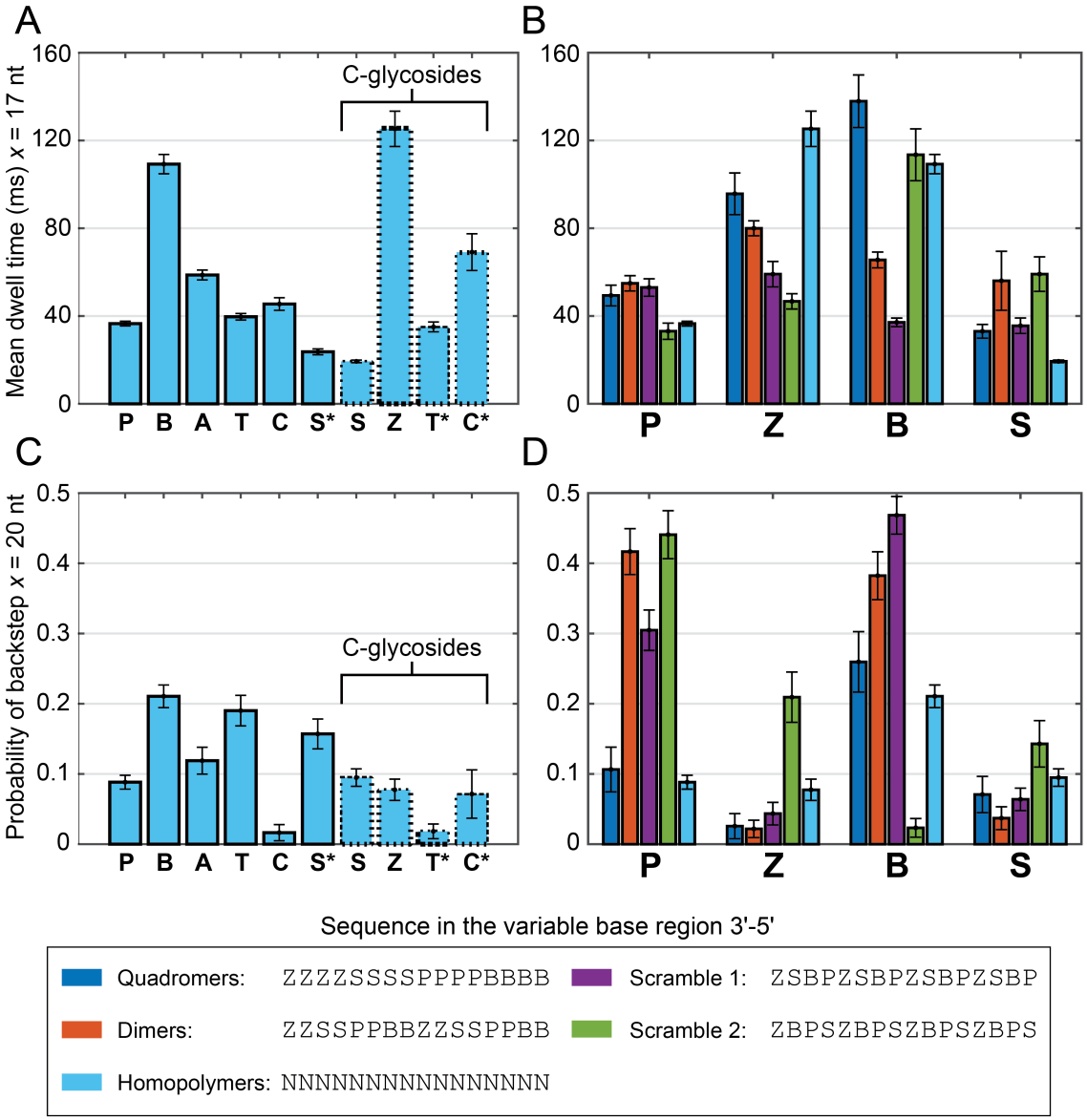

Figure 3.7: *Hachimoji* base elicit sequence specific kinetics in Hel308. (A) Mean dwell times of Hel308 translocation over homopolymers of the various tested bases. (B) Means of aggregated dwell times of Hel308 translocation over each non-standard *hachimoji* base in a template. (C) Probability of a backwards step taken by Hel308 over homopolymers of the various tested bases. (D) Aggregated probability of a backwards step taken by Hel308 over each non-standard *hachimoji* base in a template. Mean dwell time was registered to sequence at 17 nt upstream from the pore constriction, while probability of backwards step was registered to sequence at 20 nt upstream from the pore constriction. Base analogues are labeled with *. All analysis was done on the [ATP]-independent step of Hel308. All error bars are SEM.

3.6 Discussion and conclusions

These data show that the MspA nanopore system can detect and accurately distinguish all eight *hachimoji* bases and that the nanopore conductance signal from *hachimoji* homopolymers occupies a larger dynamic range of conductance than the standard DNA alphabet. This increased dynamic range means that additional information can be gleaned from the nanopore signal despite the added difficulties of sequencing 8-letter DNA. In particular, the high currents produced by **Z** and low currents produced by **P** suggest that *de novo* sequencing of a 6-letter alphabet including A, C, G, T, **P**, and **Z** is a tractable problem.

Further, by distinguishing *hachimoji* single-base substitutions with high confidence, these data show the feasibility of simultaneous direct detection of *hachimoji* nucleotides sparsely distributed within natural DNA. These are exactly the kinds of sequences generated by laboratory *in vitro* evolution (LIVE) applied to *hachimoji* libraries. Here, the binding specificity and enzyme catalytic activity of aptamers and aptazymes can be greatly increased compared to standard NAs through the addition of only a few non-standard *hachimoji* nucleotides[72, 23, 22, 14, 25]. This is one of many current and near-future applications of *hachimoji* NAs with sparse incorporation of non-standard *hachimoji* nucleotides.

Natural DNA includes additional non-canonical bases such as methylated nucleotides or lesions such as abasic residues which also manifest as ion current shifts in nanopore sequencing data. However, for many applications, only a limited set of possible sequences need to be considered, e.g., a particularly prevalent base substitution or an abasic residue which results from a particular DNA repair pathway[30]. That the nanopore system shown here can distinguish among 8 separate hypotheses with high confidence suggests that it can be readily applied to

probe *hachimoji* DNA extracted from live cells.

Denser patterns of *hachimoji* nucleotides will increase the combinatorial complexity of nanopore signals and will likely require new innovations to extract more information from the NA sequence. These include variable-voltage sequencing, frequency-domain analysis[73], as well as future work to incorporate motor enzyme kinetics as an additional independent signal into base-calling algorithms. For example, Hel308 shows strong sequence-dependent kinetics in response to *hachimoji* bases (Figure 3.7). As noncovalent nanopore-target interactions may affect the measured signal, one possibility is to design nanopores that interact with specific bases (such as the nitro group on Z), thus enhancing the current response in a base specific manner[74]. Sequencing and comparison of unmodified and transliterated NAs could also be used to glean additional information from the DNA[38].

By tracking the position along the strand where Hel308 dissociates, we found that Hel308 is partially compatible with *hachimoji* nucleotides, translocating without issue on **P** and **B** bases while retaining reasonable processivity on sparsely distributed **S** and/or **Z** bases. We also found compelling evidence that the glycosidic character of nucleosides is a major predictor of premature Hel308 strand dissociation, with C-glycosides promoting more dissociation relative to N-glycosides. We speculated this could be due to the C-glycosides adopting a different sugar pucker[75, 76], a phenomenon in which the conformation of the deoxyribose sugar switches from C2‘-endo to C3‘-endo, decreasing the distance between neighboring phosphate groups along the DNA backbone[77] (Figure 3.8). This may also be an underlying mechanism by which Hel308 discriminates between DNA and RNA in vivo as DNA and RNA exhibit C2‘-endo to C3‘-endo pucker, respectively[78].

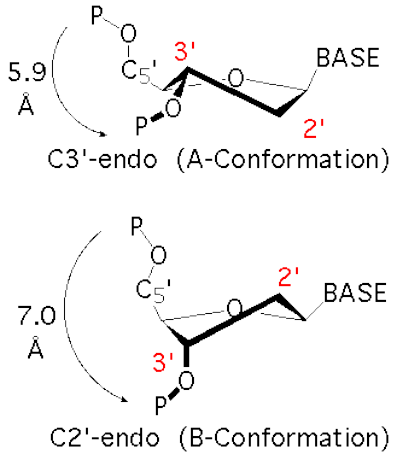

Figure 3.8: Diagram of the difference in sugar pucker between A-form and B-form DNA. C3‘-endo (A-form DNA, top) sugar pucker decreases the interphosphate distance compared to C2‘-endo (B-form DNA, bottom) sugar pucker. This difference in interphosphate distance could account for increased helicase dissociation. This figure was adapted from https://x3dna.org/highlights/sugar-pucker-correlates-with-phosphorus-base-distance.

By mutating Hel308 near the position where we localize the dissociation effect, we may be able to optimize Hel308 for use in C-glycoside nanopore sequencing. While the C-glycoside effect will hinder nanopore sequencing of **S**/**Z**-rich *hachimoji* templates actuated by wild type Hel308, the stochastic Hel308 C-glycoside dissociation rate of 12%-per-nucleotide (Figure B.5) currently allows for sequencing strands with sparse **S** and **Z** composition or short sections (<20 nt) of **S** and **Z**. More work is needed to determine the mechanism of dissociation and whether other motor enzymes used for nanopore sequencing possess similar dissociation behavior. This work may also provide useful constraints in the development of future artificial genetic systems compatible with existing NA processing enzymes.

Chapter 4

ALIEN DNA SEQUENCING

4.1 Model construction

The initial success of high accuracy reference sequencing prompted us to undertake the challenge of full *de novo* sequencing of the entirely synthetic **PZBS** genetic system, formally referred to as Alternative Isoinformational Engineered (ALIEN) DNA. Despite containing no canonical bases, ALIEN DNA was found to hybridize in the familiar A and B helical forms of standard DNA[75]. In contrast to reference sequencing, *de novo* sequencing of ALIEN DNA requires a robust sequencing model parameterized by k-mers, incorporating stochastic enzyme stepping behavior, and an algorithm (e.g. Viterbi) to determine the most likely string of k-mers composing any measured read. This overall strategy for nanopore sequencing is widely used and has proven to be effective for standard DNA and RNA nanopore sequencing[79, 52, 80]. We have a choice in the size of k-mers used to parameterize a model, with the amount of training data and the size of the pore constriction constraining the useful k-mer size. Variable voltage sequencing of standard DNA used a k-mer size of 6 nucleotides motivated by the additional DNA stretching the variable-voltage induces, but this is not feasible for ALIEN sequencing due to the immense amount of data required to sample every 6-mer.

To understand the importance of k-mer size in variable voltage sequencing, we investigated the information content of each k-mer position in the previously built ACGT 6mer model. We quantified the mutual information between k-mer model states and bases that compose k-mers at each position within the k-mer (Figure

4.1). Mutual information is defined as the amount of information learned about one variable (e.g. sequence) by measuring another (e.g. nanopore signal). We used a bin-free estimator that estimates mutual information between a discrete variable (i.e. the DNA sequence) and a continuous variable (ion current signal)[81]. We did this independently for both Hel308 kinetic substates encoded in the model. We found that the mutual information of the k-mer position and signal peaks at the center of the k-mer, as expected. This finding suggests that 3-base and 4-base ALIEN k-mer models centered over k-mer positions [-2, -1, 0, 1] and [-1, 0 ,1] respectively should be able to sequence with reasonable accuracy.

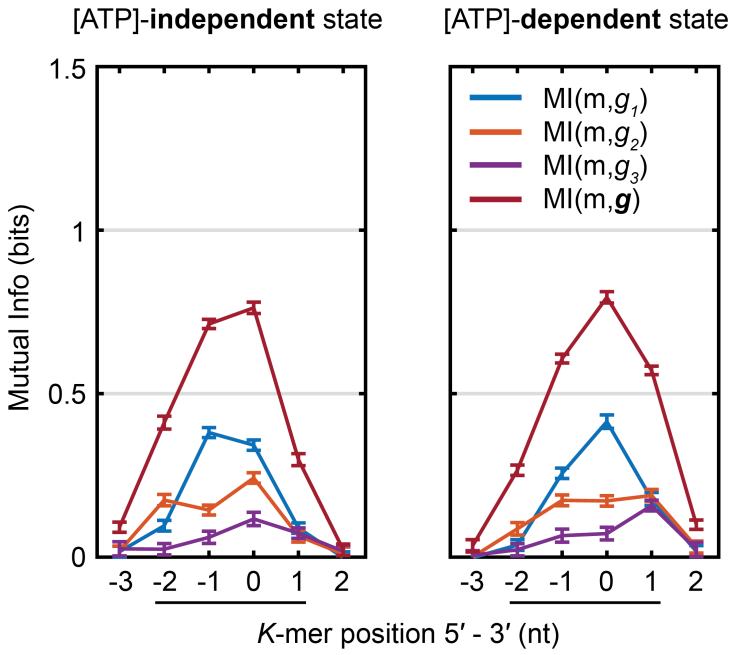

Figure 4.1: Estimation of the mutual information between the k-mer states of the ACGT 6-mer model and bases within a k-mer using a bin-free estimator[81]. K-mer position indicates the position of a base within the k-mer sequence relative to the pore constriction from 5′ to 3′. In addition to calculating the information of the entire measurement vector ($\mathbf{g}$), we separately calculate the information for each vector component (g_1, g_2, g_3) which roughly correspond to the offset, slope, and curvature of a variable voltage state. We observe that the majority of the mutual information between measurement and sequence occurs in the center of the k-mer.

Like all models, a k-mer model of nanopore conductance does not perfectly represent reality. We know that the sequence outside of any locally defined k-mer affects the measured conductance to some degree, and that a model with a larger k-mer will generally describe the data better than one with a smaller k-mer. Data limitations and model complexity limit k-mer size, since the number of unique k-mers scales exponentially with the k-mer size. Because of the perturbing effects of the sequence context outside of any k-mer, an effective k-mer model should describe the average signal of the k-mer in multiple sequence contexts as well as the variance of the average signal.

Because of the mutual information estimates suggesting that a 3- or 4-base k-mer model would be sufficient for sequencing, we chose to build an initial variable voltage ALIEN k-mer model by measuring every possible ALIEN 4-mer (256 total). We generated a 256-nucleotide long de Bruijn sequence[82] which compactly contains all such k-mers. As no genomic ALIEN DNA presently exists, we must make do with chemically synthesized oligonucleotides using phosphoramidite chemistry. Because of the limitations of phosphoramidite synthesis which include a maximum length of roughly 100 nt, this de Bruijn sequence was divided up into 11 separate oligonucleotides (Table C.1). These model-building strands were designed with a variable 28 nt region of ALIEN DNA sandwiched between two sections of ACGT DNA used for experiment calibration and helicase loading (Figure 4.2). K-mers that were not well resolved in these initial strands were placed in an additional oligonucleotide, bringing the total number of model-building strands to 12.

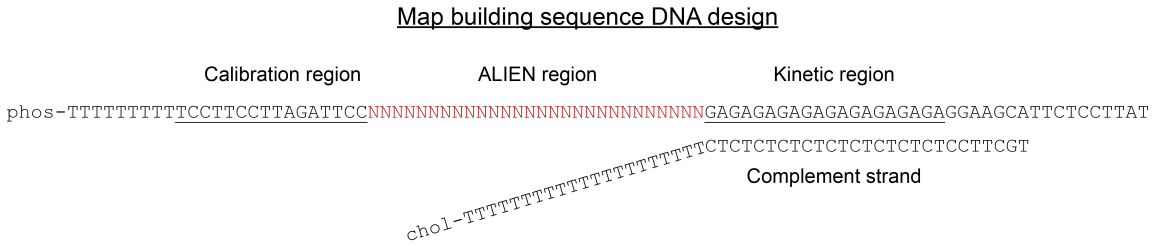

Figure 4.2: ALIEN DNA Design. We designed the ALIEN model building sequences with three key components: an ACGT calibration region, a 28 nt ALIEN region containing 25 different ALIEN k-mers per strand, and an ACGT kinetic region consisting of GA repeats and an enzyme loading region. The calibration region allows us to calibrate the nanopore signal of each experiment to all others. The ALIEN region of each sequence is a portion of the ALIEN de Bruijn sequence. The GA repeats of the kinetic region causes Hel308 to backstep at a high probability every other nucleotide when that sequence is within the helicase. This causes periodic backstepping when the ALIEN region is in the pore constriction, aiding sequence alignment to the trace.

After acquiring many reads of each model building strand, our next step was to build a consensus trace for each strand which should represent the "true" nanopore signal and omit errors present in any single read. We used a Baum-Welch expectation-maximization algorithm to align reads to a consensus, starting with an initial guess consensus and iteratively updating the consensus based on the aligned reads.

We noticed that the ALIEN reads had qualitative differences compared to reads of standard DNA. One of these differences was an increase in noise when ALIEN DNA was in the pore. This was seen within single reads as ALIEN reads begin and nominally end with standard DNA in the pore. We also observed this effect in constant voltage data from Chapter 3 (Figure B.1), suggesting that it is intrinsic to ALIEN bases and not due to any feature of the experimental setup. Interestingly among the homopolymer data, this increased noise is present only in the reads of homopolymer **B** and **Z**. This apparent base specific effect may affect the reads of the model building strands in a more uniform manner than Figure B.1 as **B** and **Z** are dispersed throughout the sequences and may influence many k-mers.

In addition to increased noise, we observed some ALIEN reads contained regions with measurements that varied significantly across reads of the same sequence. These heterogeneous regions occurred predominantly in **BS**-rich parts of the sequences, again suggesting a base-specific effect. The heterogeneity of these regions made it difficult for the consensusing process to converge on a well-defined trace.

Lastly, we also observed that ALIEN reads commonly had missing measurement states. Because we know the exact sequence of each read we measure, we expect any given read to contain a certain number of Hel308 steps excluding occasions of early enzyme fall-off. This propensity for missing states added difficulty to consensus building, as it added ambiguity as to the correct positions of consensus states. These missing states were not observed in the standard DNA regions of the reads.

With consensuses built for each strand despite the challenges outlined above, our next task was to align consensuses to each known sequence. Again the propensity for missing states in the ALIEN reads added considerable challenge. Misregistration of sequence to consensus states would occur if the consensus states were improperly arranged. Even a single missing consensus state would misregister the sequence for all states succeeding it, causing incorrect assignment of multiple k-mers. To fix misregistrations we leveraged the general current response of each base (**Z**'s give high currents, **P**'s give low currents Figure 3.2), the expected contiguity of variable voltage state curves, and cross referencing the few identical k-mers that occur in more than one strand. We were also aided by our DNA design strategy, where upstream GA repeats induced regularly spaced enzyme backstepping to help to tack down consensus state positions. After each consensus is aligned to its sequence, all that remains is to use the sequence alignment to extract and organize the consensus states by k-mer. Sequence-aligned consensuses used for model building are shown in Figures C.1 and C.2.

Three and four base k-mer models have separate advantages in our context of ALIEN sequencing. A 4-mer model has the advantage of more accurately representing the data, both from the higher level of parameterization and that its closer to the underlying physics of k-mer ion current emission. Measurements of any given 4-mer in multiple diverse sequence contexts are guaranteed to have a lower covariance than measurements of any given 3-mer, meaning the 4-mer is more constrained. The disadvantage of a 4-mer model (or any larger k-mer model) is that it requires significantly larger datasets to describe. In our model building sequences we measure the majority of 4-mers in only a single sequence context, which may not be representative of the population of a given k-mer.

Furthermore estimating the covariance of a 4-mer from multiple sequence contexts,

which is the measure of uncertainty desired for a sequencing model, is intractable when limited to a 4-mer measurement in a single sequence context. Because variable voltage measurements are multivariate (they consist of amplitudes of 3 principal components), we cannot calculate a non-singular covariance matrix for a k-mer state possessing less than three measurements (the dimensionality of the measurement). If we instead parameterize our model using a 3 base k-mer, we gain the benefit of having 4 to 7 measurements of each 3-mer in separate sequence contexts. This would allow for good covariance estimations of each k-mer.

Because of these considerations, it is unclear which k-mer model will perform best in our applications given our limited data. We opted to construct both 3 nt and 4 nt k-mer models for testing in sequencing trials. The strategy for covariance estimation differs for each model for the reasons noted above. The 3-mer model is well-sampled, so we may directly estimate the covariance from the 4 to 7 measurements of each 3-mer. To improve estimation for the undersampled 3-mers, we implemented a shrinkage method to blend a global covariance matrix of all k-mers with an individual k-mer covariance matrix to a degree dependent on the number of measurements[83] (Equation C.1). The best we can do for covariance estimates in a 4-mer model is estimate the global covariance of all k-mers and use this as a flat covariance estimate for all k-mers. We estimate this by taking differences of k-mer pairs that contain >1 measurement, then calculating the covariance of this difference (Equations C.4, C.5, C.6).

4.2 De novo single read ALIEN sequencing

To benchmark the performance of our three and four nucleotide k-mer models, we sequenced 4 strands containing 28 nt pseudorandom sequences of ALIEN DNA of

similar design to the model-building strands. We performed these experiments blind to the sequence identity of strands. Both k-mer models performed similarly, with mean single-read accuracies of 63% after unblinding (Figures 4.3A and 4.4A). This accuracy was low, as purely random basecalling with allowances for base insertions and deletions averages out to 57% (Appendix C.3). After the benchmarking trial and with knowledge of the true sequences, we added the measurements of these strands to the k-mer models. We also used the additional sequences to perform additional cross-referencing of k-mer measurements in the initial models, finding and correcting many instances of incorrect sequence registration in the model building data.

With our improved sequencing models, we repeated benchmarking with a new set of 4 blind pseudorandom ALIEN sequences. We find that sequencing using this new model results in a higher single-read accuracy than the first trial (mean accuracy = 69% 4-mer, 67% 3-mer) (Figures 4.3B and 4.4B). The breakdown of basecall errors by base shows that **B** and **S** were miscalled at a much higher rate than **P** and **Z**. That **P** and **Z** were called more accurately is not surprising, as we previously showed that measurements of **P** and **Z** are highly separated from other bases in conductance (Figure 3.2). In constrast, **B** and **S** measurements are more degenerate.

We applied the same mutual information estimator to compare the information content of initial ALIEN 4-mer model, improved ALIEN 4-mer model, and ACGT 4-mer model (Figure 4.5). We found that the improved ALIEN 4-mer model approaches the ACGT 4-mer model in information content, suggesting significant improvement over the initial ALIEN model.

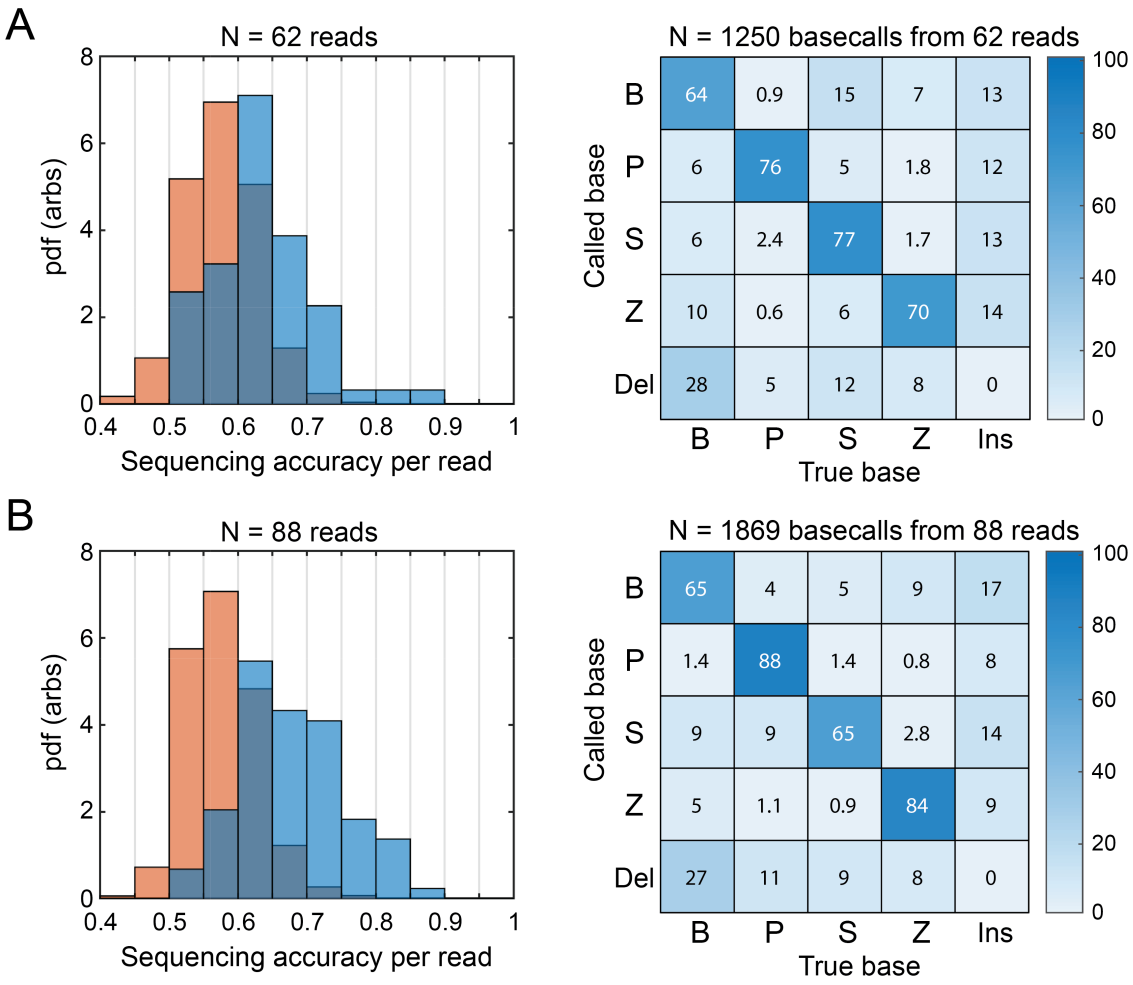

Figure 4.3: *De novo* sequencing accuracy of ALIEN DNA - 3-mer model. (A) Single-read variable-voltage sequencing accuracy and basecall confusion matrix of 4 pseudorandom strands ALIEN DNA using the initial 3-base k-mer model, mean accuracy 63%. (B) Single-read variable-voltage sequencing accuracy and basecall confusion matrix of 4 new pseudorandom strands ALIEN DNA using the improved 3-base k-mer model, mean accuracy 67%. The improved model was created through adding the measurements from the first 4 pseudorandom strands to the initial k-mer model, as well as cross-referencing the k-mers from the unblinded sequences in (A) to fix sequence misregistrations present in the initial model. Sequencing accuracy evaluated on random basecalls is shown in orange with a mean accuracy of 57%.

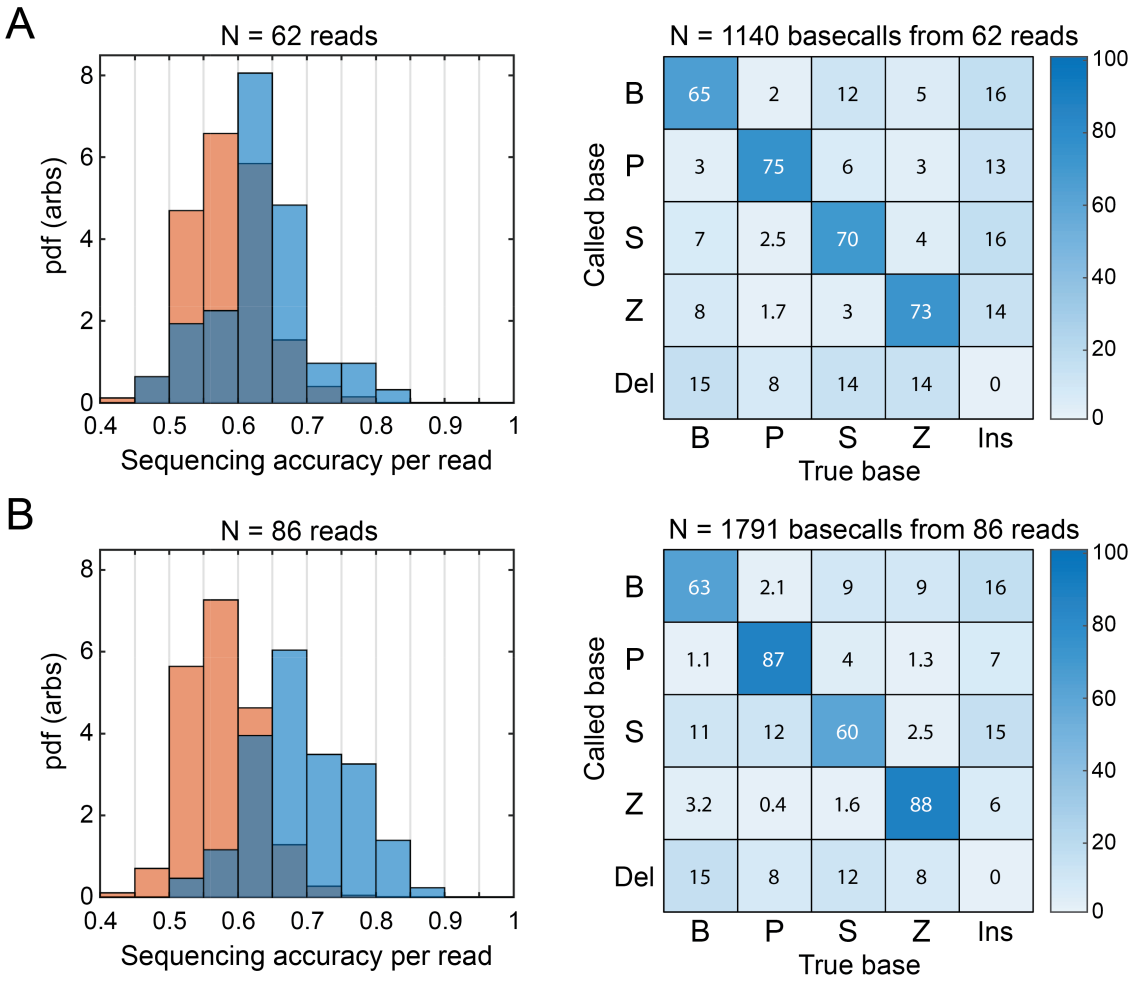

Figure 4.4: *De novo* sequencing accuracy of ALIEN DNA - 4-mer model. (A) Single-read variable-voltage sequencing accuracy and basecall confusion matrix of 4 pseudorandom strands ALIEN DNA using the initial 4-base k-mer model, mean accuracy 63%. (B) Single-read variable-voltage sequencing accuracy and basecall confusion matrix of 4 new pseudorandom strands ALIEN DNA using the improved 4-base k-mer model, mean accuracy 69%. The improved model was created through adding the measurements from the first 4 pseudorandom strands to the initial k-mer model, as well as using cross-referencing of the k-mers from the unblinded sequences in (A) to fix sequence misregistrations present in the initial model. Sequencing accuracy evaluated on random basecalls is shown in orange with a mean accuracy of 57%.

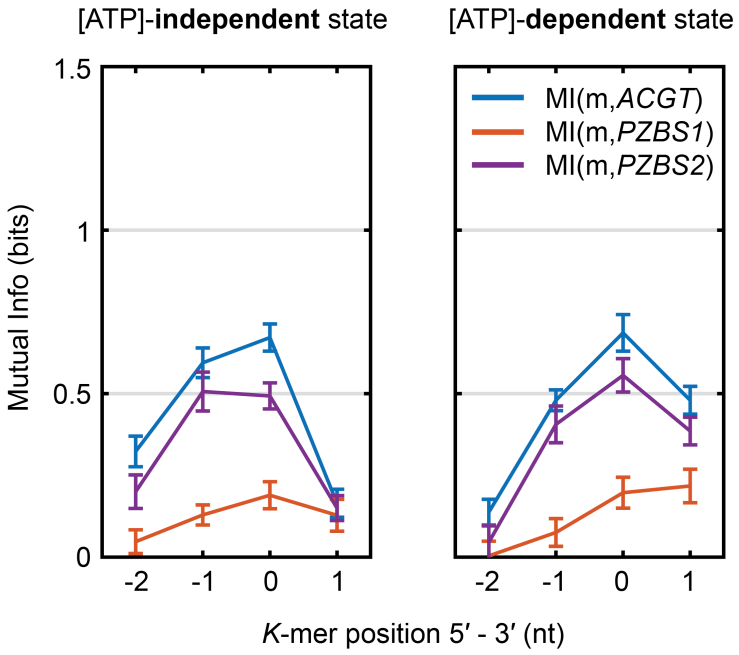

Figure 4.5: Comparison of the mutual information estimates between 4-base k-mer models of ACGT, initial ALIEN (PZBS1), and improved ALIEN (PZBS2). K-mer position indicates the position of a base within the k-mer sequence relative to the pore constriction from $5'$ to $3'$. We observe that the information content of the improved ALIEN 4-mer model approaches the ACGT 4-mer model. This suggests that our incorporation of the trial 1 data both in direct addition of measurements and fixing initial model data misregistrations significantly improved the k-mer model. Estimates were calculated using the same bin-estimator as in Figure 4.1.

4.3 Discussion and conclusions

We achieved our goal of *de novo* sequencing of ALIEN, albeit with low single-read accuracies. This is in part because our k-mer models are quite data limited compared to models of standard DNA. Because data collection was done with a small number of short oligonucleotides, our initial dataset was limited to measurements of 1-2 sequence contexts per 4-mer. For reference, our ACGT DNA k-mer model built in part through measuring kilobase scale genomic DNA contained a measurement average of 40 sequence contexts per 6-mer, enabling high-accuracy ACGT sequencing[52]. As additional ALIEN data is collected, we that error rates will continue to drop. We suspect that once we have enough data to populate the 4-mer model to the same degree as our current 3-mer model (>4 measurements per k-mer), the 4-mer model will outperform the 3-mer model given the same underlying datasets.

The same data quality issues discussed above in the context of model construction also likely contribute to the sequencing error rate. The lower accuracy of **B** and **S** basecalls may be in part due to the read heterogeneity observed in **BS**-rich regions. The prevalence of missing states in reads also contributes to the error rate, as it reduces the effective coverage of reads.

B is known to have alternate tautomeric[1] forms [84]. Free energy calculations estimate that 80% of **B**s are present as the keto tautomer, while 20% are present as undesired the enol tautomer. Since we know that nanopore signals are sensitive to minute chemical differences, the tautomerism of **B** may explain the observed signal heterogeneity of **BS**-rich regions. Depending on the timescale of tautomer interconversion, this effect may also manifest as increased noise in a nanopore signal. More work is needed to investigate the effect of tautomerization on nanopore signals.

[1]structural isomers that readily interconvert

Similar to tautomerization, **Z** may occur in different protonation states in our experiments. The 5′ nitrogen on **Z** is known to have a pKa of 7.8[25]. Since our experiments were all performed at pH 8, roughly half of the molecules of **Z** protonated at any given time. The unprotonated form of **Z** has a net negative charge, meaning that a time average of **Z** at pH 8 would result in Z having an extra half electron charge in our experiments. This means that our analyte DNA strands containing **Z** are no longer uniformly charged. This may result in unexpected stretching dynamics of the DNA in the pore due the difference in electric force felt by nucleotides of **Z**, manifesting in our signal as shifted ion current levels adjacent to the **Z** positiopn. We are currently investigating this effect through pH titration experiments.

As we discussed in Chapter 3, solid-phase synthesis of oligonucleotides introduces errors such as chemical adducts and nucleotide deletions. Little is known about the synthesis error modes of non-standard bases. Single-molecule techniques such as nanopore sequencing are sensitive to these errors and care must be taken to exclude them from analysis. We discard reads that differ from the average population, but this requires that the average population be well defined. This is not the case in the heterogeneous **BS** regions, so it is conceivable that a particularly prevalent synthesis error pathway is responsible. More investigation using highly sensitive techniques such as mass spectrometry is needed to understand potential synthesis error modes.

Our current nanopore sequencing technique is inherently low throughput, as one technician acquires only one pore and typically measures only one DNA sequence in an experiment. This is in contrast to ONT devices that are capable of running hundreds of pores in parallel. In addition no genomes exist for ALIEN DNA, meaning our data collection is limited to synthetic oligonucleotides which must be under $\sim$100 nt in length. Improving our data throughput will require parallelizing our experimental setup by allowing multiple pores to be run at once. Additionally, we may increase

throughput by pooling many sequences together in a single experiment and using standard base barcodes to deconvolute the data during analysis. It took 1.5 years to gather the data required to build the initial k-mer models for 4-letter ALIEN DNA; if we are to repeat this process for other artificial alphabets, especially ones with >4 letters, we will need to improve our data collection throughput by at least an order of magnitude. The glycoside dissociation mechanism of Hel308 also harms throughput by stochastically reducing read length. This may be improved through mutating Hel308, finding an intrinsically C-glycoside compatible motor enzyme, or using N-glycoside forms of non-standard bases.

We have demonstrated *de novo* nanopore sequencing of an entirely synthetic alphabet, but there is still work ahead to perfect this technology. We are developing tools for consensus sequencing that simultaneously incorporate read signal and basecall, combining the advantages of the two consensus strategies described above. This, along with continuing to add additional data to the k-mer models, will enable us to apply this technology to solve problems in synthetic biology. We are already planning experiments to evaluate mutant polymerase fidelity on enzymatically produced ALIEN DNA. We are also looking towards expanding *de novo* sequencing to supernumerary DNA alphabets. We have already built sequencing models for half of the *hachimoji* system when we consider both our ALIEN and standard base k-mer models. Alternatively since many applications of expanded alphabets involve only sparse incorporations of non-standard bases, we may simplify the combinatorial difficulties of large k-mer models by building models that omit k-mers of consecutive non-standard bases. With this strategy, developing effective sequencing models of even 12-letter AEGIS DNA may be tractable.

Chapter 5
CONCLUSIONS

Much progress has been made towards *de novo* sequencing of *hachimoji* DNA. In Chapter 2 I summarized all of our innovations from the last decade of nanopore sequencing that will be leveraged to sequence artificial genetic systems. In Chapter 3 I characterized *hachimoji* DNA on the nanopore and sequenced *hachimoji* single base substitutions. I also investigated Hel308 motor enzyme response to artificial bases and found not only general compatibility and sequence-specific kinetics, but a distinct chemical mechanism for enzyme fall-off from artificial DNA. In Chapter 4 I demonstrated full factorial *de novo* sequencing of an entirely synthetic DNA alphabet.

While many challenges remain before realizing the full potential of this technique, I am optimistic about the next steps. We can improve our sequencing accuracy through the integration of other parallel advancements in nanopore technology. As we have seen that artificial bases elicit sequence-specific kinetic responses in Hel308, incorporating kinetic measurements into the sequencer would boost accuracies. This is an active area of development for our group.

Our technique is inherently less accessible to the scientific community compared with the commercially-available ONT devices. It requires specialized institutional expertise and expensive instrumentation, and to date no other lab has performed MspA variable-voltage nanopore sequencing. While we have achieved the ability to sequence artificial DNA, this has much less impact if we are the only ones who can do it. To address the issues of data throughput and accessibility of the technique, I collaborated with the Marchand group at UW to synthesize 12-letter

ACGT**PZBSKXJV** DNA and sequence it with the ONT MinIon device [85]. Using large k-mer models enabled by the high throughput data collection of the MinION, we sequenced sparsely incorporated non-standard bases with reasonable sequencing accuracy. We also made available our sequencing code base, enabling any user with a MinION device to make use of our k-mer models. While MspA sequencing may be capable of higher accuracy when sequencing artificial systems, the ONT sequencing approach will almost certainly be more impactful.

Additionally, the sequencing technology developed here for artificial genetic systems may be applied to biological DNA or RNA modifications. Mapping the modifications of mRNA are of great interest to the scientific community, with their role in gene regulation being both simultaneously of great importance and poorly understood[86]. Modeling the 10 most common mRNA mods (or a subset) along with the canonical bases as an expanded genetic alphabet would allow the direct application of techniques developed in this work. Simplifying the task is the observation that mRNA mods are sparsely distributed and unlikely to occur consecutively, meaning that full factorial k-mer models are unnecessary.

The progress detailed in this dissertation hints at what the future holds for artificial nucleic acid technology. As sequencing of artificial alphabets becomes more accurate and accessible, more of the applications of these alphabets will become feasible. The next decade of synthetic nucleic acid technology holds much promise.

BIBLIOGRAPHY

[1] Andrew Travers and Georgi Muskhelishvili. DNA structure and function. *The FEBS Journal*, 282(12):2279–2295, June 2015. Publisher: John Wiley & Sons, Ltd.

[2] Paul G. Higgs and Niles Lehman. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16(1):7–17, January 2015.

[3] Nilesh B Karalkar and Steven A Benner. The challenge of synthetic biology. Synthetic Darwinism and the aperiodic crystal structure. *Synthetic Biology / Synthetic Biomolecules*, 46:188–195, October 2018.

[4] Erwin Schrödinger. *What is life? The physical aspect of the living cell and mind*. Cambridge university press Cambridge, 1944.

[5] Michael W. Grome and Farren J. Isaacs. ZTCG: Viruses expand the genetic alphabet. *Science*, 372(6541):460–461, April 2021. Publisher: American Association for the Advancement of Science.

[6] Yasuhiro Oba, Yoshinori Takano, Yoshihiro Furukawa, Toshiki Koga, Daniel P. Glavin, Jason P. Dworkin, and Hiroshi Naraoka. Identifying the wide diversity of extraterrestrial purine and pyrimidine nucleobases in carbonaceous meteorites. *Nature Communications*, 13(1):2008, April 2022.

[7] C. Ronald Geyer, Thomas R. Battersby, and Steven A. Benner. Nucleobase pairing in expanded Watson-Crick-like genetic information systems. *Structure*, 11(12):1485–1498, 2003. ISBN: 0969-2126 Publisher: Elsevier.

[8] Eric T. Kool, Juan C. Morales, and Kevin M. Guckian. Mimicking the Structure and Function of DNA: Insights into DNA Stability and Replication. *Angewandte Chemie International Edition*, 39(6):990–1009, March 2000. Publisher: John Wiley & Sons, Ltd.

[9] Karin Betz, Denis A. Malyshev, Thomas Lavergne, Wolfram Welte, Kay Diederichs, Floyd E. Romesberg, and Andreas Marx. Structural Insights into

DNA Replication without Hydrogen Bonds. *Journal of the American Chemical Society*, 135(49):18637–18643, December 2013. Publisher: American Chemical Society.

[10] Ichiro Hirao. Unnatural base pair systems for DNA/RNA-based biotechnology. *Model systems / Biopolymers*, 10(6):622–627, December 2006.

[11] Ichiro Hirao, Michiko Kimoto, Tsuneo Mitsui, Tsuyoshi Fujiwara, Rie Kawai, Akira Sato, Yoko Harada, and Shigeyuki Yokoyama. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. *Nature Methods*, 3(9):729–735, September 2006.

[12] Omid Khakshoor, Steven E. Wheeler, K. N. Houk, and Eric T. Kool. Measurement and Theory of Hydrogen Bonding Contribution to Isosteric DNA Base Pairs. *Journal of the American Chemical Society*, 134(6):3154–3163, February 2012. Publisher: American Chemical Society.

[13] Zunyi Yang, Daniel Hutter, Pinpin Sheng, A. Michael Sismour, and Steven A. Benner. Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. *Nucleic Acids Research*, 34(21):6095–6101, November 2006.

[14] Elisa Biondi and Steven A. Benner. Artificially Expanded Genetic Information Systems for New Aptamer Technologies. *Biomedicines*, 6(2):53, 2018.

[15] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, January 2012.

[16] Javier T. Granados-Riveron and Guillermo Aquino-Jarquin. Engineering of the current nucleoside-modified mRNA-LNP vaccines against SARS-CoV-2. *Biomedicine & Pharmacotherapy*, 142:111953, October 2021.

[17] Zunyi Yang and Steven A Benner. Compositions for the Multiplexed Detection of Viruses. US Patent 20220162600-A1, November 2020.

[18] Lyudmyla G. Glushakova, Andrea Bradley, Kevin M. Bradley, Barry W. Alto, Shuichi Hoshika, Daniel Hutter, Nidhi Sharma, Zunyi Yang, Myong-Jung Kim, and Steven A. Benner. High-throughput multiplexed xMAP Luminex array panel for detection of twenty two medically important mosquito-borne arboviruses

based on innovations in synthetic biology. *Journal of Virological Methods*, 214:60–74, March 2015.

[19] Shuichi Hoshika, Fei Chen, Nicole A. Leal, and Steven A. Benner. Artificial Genetic Systems: Self-Avoiding DNA in PCR and Multiplexed PCR. *Angewandte Chemie International Edition*, 49(32):5554–5557, July 2010. Publisher: John Wiley & Sons, Ltd.

[20] Zunyi Yang, Fei Chen, Stephen G. Chamberlin, and Steven A. Benner. Expanded Genetic Alphabets in the Polymerase Chain Reaction. *Angewandte Chemie International Edition*, 49(1):177–180, January 2010. Publisher: John Wiley & Sons, Ltd.

[21] Regina Stoltenburg, Christine Reinemann, and Beate Strehlitz. SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering*, 24(4):381–403, October 2007.

[22] Liqin Zhang, Zunyi Yang, Thu Le Trinh, I-Ting Teng, Sai Wang, Kevin M. Bradley, Shuichi Hoshika, Qunfeng Wu, Sena Cansiz, Diane J. Rowold, Christopher McLendon, Myong-Sang Kim, Yuan Wu, Cheng Cui, Yuan Liu, Weijia Hou, Kimberly Stewart, Shuo Wan, Chen Liu, Steven A. Benner, and Weihong Tan. Aptamers against Cells Overexpressing Glypican 3 from Expanded Genetic Systems Combined with Cell Engineering and Laboratory Evolution. *Angewandte Chemie International Edition*, 55(40):12372–12375, September 2016. Publisher: John Wiley & Sons, Ltd.

[23] Elisa Biondi, Joshua D. Lane, Debasis Das, Saurja Dasgupta, Joseph A. Piccirilli, Shuichi Hoshika, Kevin M. Bradley, Bryan A. Krantz, and Steven A. Benner. Laboratory evolution of artificially expanded DNA gives redesignable aptamers that target the toxic form of anthrax protective antigen. *Nucleic Acids Research*, 44(20):9565–9577, November 2016.

[24] Liqin Zhang, Sai Wang, Zunyi Yang, Shuichi Hoshika, Sitao Xie, Jin Li, Xigao Chen, Shuo Wan, Long Li, Steven A. Benner, and Weihong Tan. An Aptamer-Nanotrain Assembled from Six-Letter DNA Delivers Doxorubicin Selectively to Liver Cancer Cells. *Angewandte Chemie International Edition*, 59(2):663–668, January 2020. Publisher: John Wiley & Sons, Ltd.

[25] Craig A. Jerome, Shuichi Hoshika, Kevin M. Bradley, Steven A. Benner, and Elisa Biondi. In vitro evolution of ribonucleases from expanded genetic alphabets.

Proceedings of the National Academy of Sciences, 119(44):e2208261119, November 2022. Publisher: Proceedings of the National Academy of Sciences.

[26] Liqin Zhang, Zunyi Yang, Kwame Sefah, Kevin M. Bradley, Shuichi Hoshika, Myong-Jung Kim, Hyo-Joong Kim, Guizhi Zhu, Elizabeth Jiménez, Sena Cansiz, I-Ting Teng, Carole Champanhac, Christopher McLendon, Chen Liu, Wen Zhang, Dietlind L. Gerloff, Zhen Huang, Weihong Tan, and Steven A. Benner. Evolution of Functional Six-Nucleotide DNA. *Journal of the American Chemical Society*, 137(21):6734–6737, June 2015. Publisher: American Chemical Society.

[27] Leping Sun, Xingyun Ma, Binliang Zhang, Yanjia Qin, Jiezhao Ma, Yuhui Du, and Tingjian Chen. From polymerase engineering to semi-synthetic life: artificial expansion of the central dogma. *RSC Chemical Biology*, 3(10):1173–1197, 2022. Publisher: Royal Society of Chemistry.

[28] Denis A. Malyshev, Kirandeep Dhami, Thomas Lavergne, Tingjian Chen, Nan Dai, Jeremy M. Foster, Ivan R. Corrêa, and Floyd E. Romesberg. A semi-synthetic organism with an expanded genetic alphabet. *Nature*, 509(7500):385–388, May 2014.

[29] Yorke Zhang, Brian M. Lamb, Aaron W. Feldman, Anne Xiaozhou Zhou, Thomas Lavergne, Lingjun Li, and Floyd E. Romesberg. A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proceedings of the National Academy of Sciences*, 114(6):1317–1322, February 2017. Publisher: Proceedings of the National Academy of Sciences.

[30] Michael P. Ledbetter, Jonathan M. Craig, Rebekah J. Karadeema, Matthew T. Noakes, Hwanhee C. Kim, Sarah J. Abell, Jesse R. Huang, Brooke A. Anderson, Ramanarayanan Krishnamurthy, Jens H. Gundlach, and Floyd E. Romesberg. Nanopore Sequencing of an Expanded Genetic Alphabet Reveals High-Fidelity Replication of a Predominantly Hydrophobic Unnatural Base Pair. *Journal of the American Chemical Society*, 142(5):2110–2114, February 2020. Publisher: American Chemical Society.

[31] Yorke Zhang, Jerod L. Ptacin, Emil C. Fischer, Hans R. Aerni, Carolina E. Caffaro, Kristine San Jose, Aaron W. Feldman, Court R. Turner, and Floyd E. Romesberg. A semi-synthetic organism that stores and retrieves increased genetic information. *Nature*, 551(7682):644–647, November 2017.

[32] Luis Ceze, Jeff Nivala, and Karin Strauss. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8):456–466, August 2019.

[33] Steven A. Benner. Understanding Nucleic Acids Using Synthetic Chemistry. *Accounts of Chemical Research*, 37(10):784–797, October 2004. Publisher: American Chemical Society.

[34] Hyo-Joong Kim and Steven A. Benner. Prebiotic stereoselective synthesis of purine and noncanonical pyrimidine nucleotide from nucleobases and phosphorylated carbohydrates. *Proceedings of the National Academy of Sciences*, 114(43):11315–11320, October 2017. Publisher: Proceedings of the National Academy of Sciences.

[35] Frederick Sanger, Steven Nicklen, and Alan R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, December 1977. ISBN: 0027-8424 Publisher: National Acad Sciences.

[36] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

[37] Evelien M Bunnik and Karine G Le Roch. An introduction to functional genomics and systems biology. *Advances in wound care*, 2(9):490–498, 2013. Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

[38] Zunyi Yang, Fei Chen, J. Brian Alvarado, and Steven A. Benner. Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. *Journal of the American Chemical Society*, 133(38):15105–15112, September 2011. Publisher: American Chemical Society.

[39] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, 30(4):349–353, April 2012.

[40] Andrew H Laszlo, Ian M Derrington, Brian C Ross, Henry Brinkerhoff, Andrew Adey, Ian C Nova, Jonathan M Craig, Kyle W Langford, Jenny Mae Samson, Riza Daza, Kenji Doering, Jay Shendure, and Jens H Gundlach. Decoding long

nanopore sequencing reads of natural DNA. *Nature Biotechnology*, 32(8):829–833, August 2014.

[41] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8, March 2015.

[42] Cees Dekker. Solid-state nanopores. *Nature Nanotechnology*, 2(4):209–215, April 2007.

[43] Anastassia A. Vorobieva, Paul White, Binyong Liang, Jim E. Horne, Asim K. Bera, Cameron M. Chow, Stacey Gerben, Sinduja Marx, Alex Kang, Alyssa Q. Stiving, Sophie R. Harvey, Dagan C. Marx, G. Nasir Khan, Karen G. Fleming, Vicki H. Wysocki, David J. Brockwell, Lukas K. Tamm, Sheena E. Radford, and David Baker. De novo design of transmembrane β barrels. *Science*, 371(6531):eabc8182, February 2021. Publisher: American Association for the Advancement of Science.

[44] John J. Kasianowicz, Eric Brandin, Daniel Branton, and David W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, November 1996. Publisher: Proceedings of the National Academy of Sciences.

[45] Stefan Howorka and Hagan Bayley. Probing distance and electrical potential within a protein pore with tethered DNA. *Biophysical Journal*, 83(6):3202–3210, 2002. ISBN: 0006-3495 Publisher: Elsevier.

[46] Michael Faller, Michael Niederweis, and Georg E. Schulz. The structure of a mycobacterial outer-membrane channel. *Science*, 303(5661):1189–1192, 2004. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[47] Tom Z. Butler, Mikhail Pavlenok, Ian M. Derrington, Michael Niederweis, and Jens H. Gundlach. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proceedings of the National Academy of Sciences*, 105(52):20647–20652, December 2008.

[48] Andrew H. Laszlo, Ian M. Derrington, and Jens H. Gundlach. MspA nanopore as a single-molecule tool: From sequencing to SPRNT. *Single molecule probing by fluorescence and force detection*, 105:75–89, August 2016.

[49] Christopher A. Thomas, Jonathan M. Craig, Shuichi Hoshika, Henry Brinkerhoff, Jesse R. Huang, Sarah J. Abell, Hwanhee C. Kim, Michaela C. Franzi, Jessica D. Carrasco, Hyo-Joong Kim, Drew C. Smith, Jens H. Gundlach, Steven A. Benner, and Andrew H. Laszlo. Assessing Readability of an 8-Letter Expanded Deoxyribonucleic Acid Alphabet with Nanopores. *Journal of the American Chemical Society*, 145(15):8560–8568, April 2023. Publisher: American Chemical Society.

[50] Petersen Lauren M., Martin Isabella W., Moschetti Wayne E., Kershaw Colleen M., and Tsongalis Gregory J. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. *Journal of Clinical Microbiology*, 58(1):10.1128/jcm.01315–19, December 2019. Publisher: American Society for Microbiology.

[51] Ian M Derrington, Jonathan M Craig, Eric Stava, Andrew H Laszlo, Brian C Ross, Henry Brinkerhoff, Ian C Nova, Kenji Doering, Benjamin I Tickman, Mostafa Ronaghi, Jeffrey G Mandell, Kevin L Gunderson, and Jens H Gundlach. Subangstrom single-molecule measurements of motor proteins using a nanopore. *Nature Biotechnology*, 33(10):1073–1075, October 2015.

[52] Matthew T. Noakes, Henry Brinkerhoff, Andrew H. Laszlo, Ian M. Derrington, Kyle W. Langford, Jonathan W. Mount, Jasmine L. Bowman, Katherine S. Baker, Kenji M. Doering, Benjamin I. Tickman, and Jens H. Gundlach. Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nature Biotechnology*, 37(6):651–656, June 2019.

[53] Sepideh Tavakoli, Mohammad Nabizadeh, Amr Makhamreh, Howard Gamper, Caroline A. McCormick, Neda K. Rezapour, Ya-Ming Hou, Meni Wanunu, and Sara H. Rouhanifard. Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nature Communications*, 14(1):334, January 2023.

[54] Jonathan M. Craig, Andrew H. Laszlo, Henry Brinkerhoff, Ian M. Derrington, Matthew T. Noakes, Ian C. Nova, Benjamin I. Tickman, Kenji Doering, Noah F. de Leeuw, and Jens H. Gundlach. Revealing dynamics of helicase translocation on single-stranded DNA using high-resolution nanopore tweezers. *Proceedings of the National Academy of Sciences*, 114(45):11932–11937, November 2017.

[55] Jonathan M Craig, Andrew H Laszlo, Ian C Nova, Henry Brinkerhoff, Matthew T Noakes, Katherine S Baker, Jasmine L Bowman, Hugh R Higinbotham,

Jonathan W Mount, and Jens H Gundlach. Determining the effects of DNA sequence on Hel308 helicase translocation along single-stranded DNA using nanopore tweezers. *Nucleic Acids Research*, 47(5):2506–2513, January 2019.

[56] Jonathan M. Craig, Andrew H. Laszlo, Ian C. Nova, and Jens H. Gundlach. Modelling single-molecule kinetics of helicase translocation using high-resolution nanopore tweezers (SPRNT). *Essays in Biochemistry*, 65(1):109–127, April 2021.

[57] Andrew H. Laszlo, Jonathan M. Craig, Momčilo Gavrilov, Ramreddy Tippana, Ian C. Nova, Jesse R. Huang, Hwanhee C. Kim, Sarah J. Abell, Mallory deCampos Stairiker, Jonathan W. Mount, Jasmine L. Bowman, Katherine S. Baker, Hugh Higinbotham, Dmitriy Bobrovnikov, Taekjip Ha, and Jens H. Gundlach. Sequence-dependent mechanochemical coupling of helicase translocation and unwinding at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 119(36):e2202489119, September 2022. Publisher: Proceedings of the National Academy of Sciences.

[58] Jonathan M Craig, Maria Mills, Hwanhee C Kim, Jesse R Huang, Sarah J Abell, Jonathan W Mount, Jens H Gundlach, Keir C Neuman, and Andrew H Laszlo. Nanopore tweezers measurements of RecQ conformational changes reveal the energy landscape of helicase motion. *Nucleic Acids Research*, 50(18):10601–10613, October 2022.

[59] Joseph H. Chapman, Jonathan M. Craig, Clara D. Wang, Jens H. Gundlach, Keir C. Neuman, and J. Robert Hogg. UPF1 mutants with intact ATPase but deficient helicase activities promote efficient nonsense-mediated mRNA decay. *Nucleic Acids Research*, 50(20):11876–11894, 2022. ISBN: 0305-1048 Publisher: Oxford University Press.

[60] Sinduja K Marx, Keith J Mickolajczyk, Jonathan M Craig, Christopher A Thomas, Akira M Pfeffer, Sarah J Abell, Jessica D Carrasco, Michaela C Franzi, Jesse R Huang, Hwanhee C Kim, Henry Brinkerhoff, Tarun M Kapoor, Jens H Gundlach, and Andrew H Laszlo. Observing inhibition of the SARS-CoV-2 helicase at single-nucleotide resolution. *Nucleic Acids Research*, 51(17):9266–9278, September 2023.

[61] Alan Shaw, Jonathan M Craig, Hossein Amiri, Jeonghoon Kim, Heather E Upton, Sydney C Pimentel, Jesse R Huang, Susan Marqusee, Kathleen Collins, Jens H Gundlach, and Carlos J Bustamante. Nanopore molecular trajectories of a eukaryotic reverse transcriptase reveal a long-range RNA structure sensing

mechanism. *bioRxiv : the preprint server for biology*, page 2023.04.05.535757, November 2023.

[62] Jonathan M. Craig, Andrew H. Laszlo, Ian M. Derrington, Brian C. Ross, Henry Brinkerhoff, Ian C. Nova, Kenji Doering, Benjamin I. Tickman, Mark T. Svet, and Jens H. Gundlach. Direct Detection of Unnatural DNA Nucleotides dNaM and d5SICS using the MspA Nanopore. *PLOS ONE*, 10(11):e0143253, November 2015. Publisher: Public Library of Science.

[63] Andrew H. Laszlo, Ian M. Derrington, Henry Brinkerhoff, Kyle W. Langford, Ian C. Nova, Jenny Mae Samson, Joshua J. Bartlett, Mikhail Pavlenok, and Jens H. Gundlach. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences*, 110(47):18904–18909, November 2013.

[64] Jared T Simpson, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, April 2017.

[65] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, November 2021.

[66] Henry Brinkerhoff, Albert S. W. Kang, Jingqian Liu, Aleksei Aksimentiev, and Cees Dekker. Multiple rereads of single proteins at single–amino acid resolution using nanopores. *Science*, 374(6574):1509–1513, December 2021. Publisher: American Association for the Advancement of Science.

[67] Frank Seela, Changfu Wei, and Alexander Melenewski. Isoguanine Quartets Formed by d(T4isoG4T4): Tetraplex Identification and Stability. *Nucleic Acids Research*, 24(24):4940–4945, December 1996.

[68] John C. Chaput and Christopher Switzer. A DNA pentaplex incorporating nucleobase quintets. *Proceedings of the National Academy of Sciences*, 96(19):10614–10619, September 1999. Publisher: Proceedings of the National Academy of Sciences.

[69] Sergey M. Bezrukov and John J. Kasianowicz. Current noise reveals protonation kinetics and number of ionizable sites in an open protein ion channel. *Physical*

Review Letters, 70(15):2352–2355, April 1993. Publisher: American Physical Society.

[70] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958. Publisher: Taylor & Francis.

[71] Katharina Büttner, Sebastian Nehring, and Karl-Peter Hopfner. Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nature Structural & Molecular Biology*, 14(7):647–652, July 2007.

[72] Shuichi Hoshika, Nicole A. Leal, Myong-Jung Kim, Myong-Sang Kim, Nilesh B. Karalkar, Hyo-Joong Kim, Alison M. Bates, Norman E. Watkins, Holly A. SantaLucia, Adam J. Meyer, Saurja DasGupta, Joseph A. Piccirilli, Andrew D. Ellington, John SantaLucia, Millie M. Georgiadis, and Steven A. Benner. Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science*, 363(6429):884–887, February 2019.

[73] Xinyi Li, Yi-Lun Ying, Xi-Xin Fu, Yong-Jing Wan, and Yi-Tao Long. Single-Molecule Frequency Fingerprint for Ion Interaction Networks in a Confined Nanopore. *Angewandte Chemie International Edition*, 60(46):24582–24587, November 2021. Publisher: John Wiley & Sons, Ltd.

[74] Meng-Yin Li, Yi-Lun Ying, Jie Yu, Shao-Chuang Liu, Ya-Qian Wang, Shuang Li, and Yi-Tao Long. Revisiting the Origin of Nanopore Current Blockage for Volume Difference Sensing at the Atomic Level. *JACS Au*, 1(7):967–976, July 2021. Publisher: American Chemical Society.

[75] Shuichi Hoshika, Madhura S. Shukla, Steven A. Benner, and Millie M. Georgiadis. Visualizing "Alternative Isoinformational Engineered" DNA in A- and B-Forms at High Resolution. *Journal of the American Chemical Society*, 144(34):15603–15611, August 2022. Publisher: American Chemical Society.

[76] Darrell R. Davis. Stabilization of RNA stacking by pseudouridine. *Nucleic Acids Research*, 23(24):5020–5026, December 1995.

[77] Wolfram Saenger, William N. Hunter, and Olga Kennard. DNA conformation is determined by economics in the hydration of phosphate groups. *Nature*, 324(6095):385–388, November 1986.

[78] Alexander Rich. The double helix: a tale of two puckers. *Nature Structural & Molecular Biology*, 10(4):247–249, April 2003.

[79] Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1):90, July 2018.

[80] Aaron M. Fleming, Justin C. Dingman, Yizhou Wu, Spencer S. Hoon, and Cynthia J. Burrows. Nanopore Direct RNA Sequencing for Modified Uridine Nucleotides Yields Signals Dependent on the Physical Properties of the Modified Base. *Israel Journal of Chemistry*, n/a(n/a):e202300177, January 2024. Publisher: John Wiley & Sons, Ltd.

[81] Brian C. Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, USA.

[82] Nicolaas Govert De Bruijn. A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 49(7):758–764, 1946.

[83] Jerome H. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[84] Lukas Eberlein, Frank R. Beierlein, Nico J. R. van Eikema Hommes, Ashish Radadiya, Jochen Heil, Steven A. Benner, Timothy Clark, Stefan M. Kast, and Nigel G. J. Richards. Tautomeric Equilibria of Nucleobases in the Hachimoji Expanded Genetic Alphabet. *Journal of Chemical Theory and Computation*, 16(4):2766–2777, April 2020. Publisher: American Chemical Society.

[85] Hinako Kawabe, Christopher A. Thomas, Shuichi Hoshika, Myong-Jung Kim, Myong-Sang Kim, Logan Miessner, Nicholas Kaplan, Jonathan M. Craig, Jens H. Gundlach, Andrew H. Laszlo, Steven A. Benner, and Jorge A. Marchand. Enzymatic synthesis and nanopore sequencing of 12-letter supernumerary DNA. *Nature Communications*, 14(1):6820, October 2023.

[86] Boxuan Simen Zhao, Ian A. Roundtree, and Chuan He. Post-transcriptional gene regulation by mRNA modifications. *Nature reviews Molecular cell biology*,

18(1):31–42, 2017. ISBN: 1471-0072 Publisher: Nature Publishing Group UK London.

[87] Colin H. LaMont and Paul A. Wiggins. The Development of an Information Criterion for Change-Point Analysis. *Neural Computation*, 28(3):594–612, March 2016.

Appendix A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

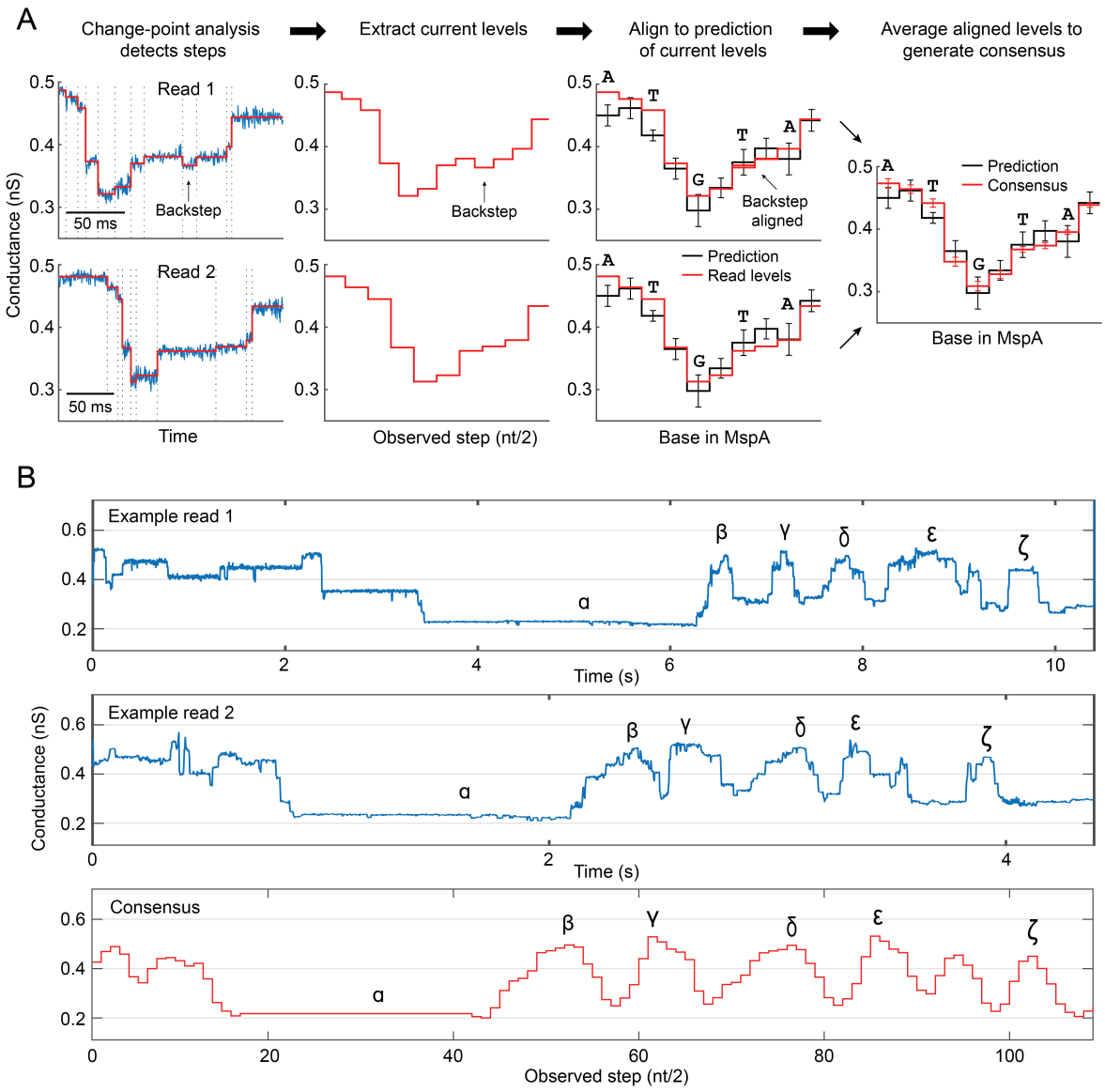

Figure A.1: Generating a consensus. (A) Raw data is segmented into levels using change-point analysis [87]. These levels are extracted omitting time information and aligned to a level prediction based on DNA sequence[52]. Averaging the aligned levels from many reads generates a consensus. Consensus errors are standard deviation. (B) (Top/Middle) Two representative nanopore traces of the *hachimoji* DNA template Poly P actuated by Hel308. (Bottom) Full consensus of the Poly P DNA template. Common conductance patterns within the traces are highlighted by Greek letters. This figure was adapted from [49].

Variable-voltage nanopore data analysis

A full description of the variable-voltage data analysis pipeline is described in [52]. To briefly summarize:

1. The raw data is segmented into events and enzyme steps via change point detection.

2. The capacitative charging current is subtracted.

3. The time-ordered conductance vs voltage curves are sampled at voltages such that the conductance is sampled uniformly over DNA position.

4. The resulting 101-dimensional conductance curves ($\boldsymbol{g}_{101}$) are normalized to remove the position-independent component of the signal.

5. We use principal component analysis to reduce the dimensionality of the conductance-voltage curves that describe each state (Figure 2.7).

Variable voltage conductance state normalization

Normalization of variable voltage conductance states is a non-linear transformation of the $\boldsymbol{g}_{101}$ data done to remove the component of the signal that is not position-dependent. This process has two components: subtracting the mean conductance of a read from every conductance state and correct for "fraying" of each conductance curve. "Fraying" refers to the observation of exaggerated signal response due to DNA stretching at high voltage, reducing the number of bases that contribute to the signal. "Fraying" is corrected by linearly fitting a conductance curve to obtain a slope, which is then itself linearly fit to the voltages of the conductance curve. This fit represents the linear voltage response as a function of conductance which can be subtracted

from the conductance curve to obtain the final normalized signal which is used for all downstream analysis. A full description of this normalization transformation is described in [52].

Appendix B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

The work discussed in this appendix was published and reproduced with permission from Thomas, CA. *et al.* Assessing Readability of an 8-Letter Expanded Deoxyribosenucleic Acid Alphabet with Nanopores. *Journal of the American Chemical Society,* 145(15):8560-8568, April 2023.[49] Copyright 2023 American Chemical Society.

B.1 Extended Materials and Methods

B.1.1 Pore Establishment

A single M2-NNN MspA nanopore was established in a 1,2-di-O-phytanyl-sn-glycero-3-phosphocholine (DOPHPC) lipid bilayer using methods that have been well established [48]. Lipids were from Avanti Polar Lipids.

B.1.2 Operating Conditions

All experiments were run at 400 mM *trans* [KCl] and 400 mM *cis* [KCl] with 10 mM HEPES at pH 8.0 and 10 mM [$MgCl_2$] (*cis* only) at temperature 37 ± 1 degrees Celsius. Once a single M2-NNN MspA nanopore was established, a buffer with the above conditions along with ATP was perfused to the *cis* well. ATP was ordered from Sigma Aldrich. The perfusion is done to maintain constant concentrations of the reactants/products in the reaction volume. DNA, DTT, and Hel308 were added to

final concentrations of 10 nM, 1 mM, and 200 nM, respectively. Reactants/products were re-perfused every 45 minutes.

B.1.3 Proteins

M2-NNN MspA (accession number CAB56052.1) were prepared as described previously [48][51]. We used Hel308 from Thermococcus gammatolerans EJ3 (accession number WP_015858487.1), which expressed using standard techniques by in-house facilities.

B.1.4 DNA preparation

AEGIS phosphoramidites (Firebird Biomolecular Sciences LLC.
www.firebirdbio.com) were used in solid phase oligonucleotide synthesis on an ABI 394 instrument to make the non-standard DNA molecules. Oligonucleotides were purified on 10% denatured PAGE and then desalted on C18 Sep-Pak cartridge. DNA strands were suspended in 100 mM KCl at 20 μM concentration. Sequences were then mixed with their complement and were annealed at 90°C and then decreased step-wise by $\sim$ 10°C per minute to 4°C. Once annealed, these sequences were diluted to 1 μM for addition to the experiment. A complete list of DNA strands used in this work is given in Table C.1).

B.1.5 Chemical names

P 2-amino-8-(1′-β-D-2′-deoxyribofuranosyl)-imidazo-[1,2a]-1,3,5-triazin-[8H]-4-one

Z 6-amino-3-(14′-β-D-2′-deoxyribofuranosyl)-5-nitro-1H-pyridin-2-one

S 3-methyl-6-amino-5-(1′-β-D-2′-deoxyribofuranosyl)-pyrimidin-2-one

B isoguanine 6-amino-9[(1′-β-D-2′-deoxyribofuranosyl)-4-hydroxy-5-(hydroxymethyl)-oxolan-2-yl]-1H-purin-2-one

***T* pseudothymine** 5-[(2S,3R,4S,5R)-3,4-Dihydroxy-5-(hydroxymethyl)oxolan-2-yl]-1-methylpyrimidine-2,4-dione

***S* isocytidine** 5-methyl-isocytosine

***C* pseudoisocytidine** 2-amino-5-[(2S,3R,4S,5R)-3,4-dihydroxy-5-(hydroxymethyl)oxolan-2-yl]-1H-pyrimidin-6-one

B.1.6 Data Acquisition

Data was acquired with custom labview software on an Axopatch 200B amplifier at 50 kHz, and downsampled by averaging to 5 kHz. In variable-voltage experiments, we applied a 200 Hz, 100 mV peak-to-peak triangle waveform in addition to a constant 150 mV DC offset. Because we filter the input voltage with a 3.5 kHz lowpass filter, the voltage waveform spans only $\sim$97 mV. In constant-voltage experiments, we applied a constant 180 mV DC offset.

Name	Sequence (5′ - 3′)
SBS A	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCAACGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS C	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCACCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS G	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCAGCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS T	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCATCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS P	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCAPCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS Z	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCAZCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS B	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCABCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
SBS S	/phos/TTTTTTTTTTTTTCTTAGATTCCAATGTCASCGTCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly A	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTTTTTTTTTTTTTTTTTTCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly T	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTAAAAAAAAAAAAAAAAACACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly C	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTCCCCCCCCCCCCCCCCCCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly P	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTPPPPPPPPPPPPPPPPPCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly Z	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTZZZZZZZZZZZZZZZZZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly B	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTBBBBBBBBBBBBBBBBBCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly S	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTSSSSSSSSSSSSSSSSSCACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly isoC	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGT$\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}\not{S}$CACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly pseudoT	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGT$\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}\not{T}$CACGGCAGACAACTCATGCAGGAAGCATTCTC
Poly pseudoisoC	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGT$\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}\not{C}$CACGGCAGACAACTCATGCAGGAAGCATTCTC
Non-standard quadromers	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTBBBBPPPPSSSSSZZZZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Non-standard dimers	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTBBPPSSZZBBPPSSZZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Non-standard scramble 1	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTPBSZPBSZPBSZPBSZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Non-standard scramble 2	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTSPBZSPBZSPBZSPBZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Standard complement	CCTGCATGAGTTGTCTGCCGTGXXXX/chol/

Table B.1: Complete list of DNA sequences used for experiments in chapter 3. /phos/ = 5′ phosphate, /chol/ = 3′ cholesterol tag, X = 18-carbon spacer. Standard complement was annealed to all strands. Base analogues are indicated by strikethroughs: $\not{S}$ = isoC, $\not{T}$ = pseudoT, and $\not{C}$ = pseudoisoC.

Name	Number of reads
SBS A	63
SBS C	198
SBS G	124
SBS T	88
SBS P	105
SBS Z	68
SBS B	52
SBS S	226
Poly A	63
Poly C	146
Poly T	101
Poly P	92
Poly Z	85
Poly B	63
Poly S	100
Poly isoC	73
Poly pseudoT	104
Poly pseudoisoC	94
Non-standard quadromers	143
Non-standard dimers	160
Non-standard scramble 1	115
Non-standard scramble 2	80

Table B.2: Summary of all nanopore reads collected for each DNA template used in chapter 3 The SBS read numbers reflect the omission of 50 reads from analysis that were used to construct the consensus for each template.

B.2 Measuring the conductances of hachimoji homopolymers

To measure homopolymer conductances, we selected several representative ion current vs time homopolymer reads and identified the homopolymer region based on its long semi-flat signal. We excised the ion current data points from these regions and divided the data by the local open pore current for each read to obtain a normalized current profile for all reads. To convert into conductance, we multiplied the normalized current profiles with the mean of the local open pore currents (pA) and divided by the applied voltage (mV) to obtain the conductance profile in nanosiemens (nS). Four representative ion current conductance traces for each homopolymer template are shown in Figure B.1 and Figure B.2.

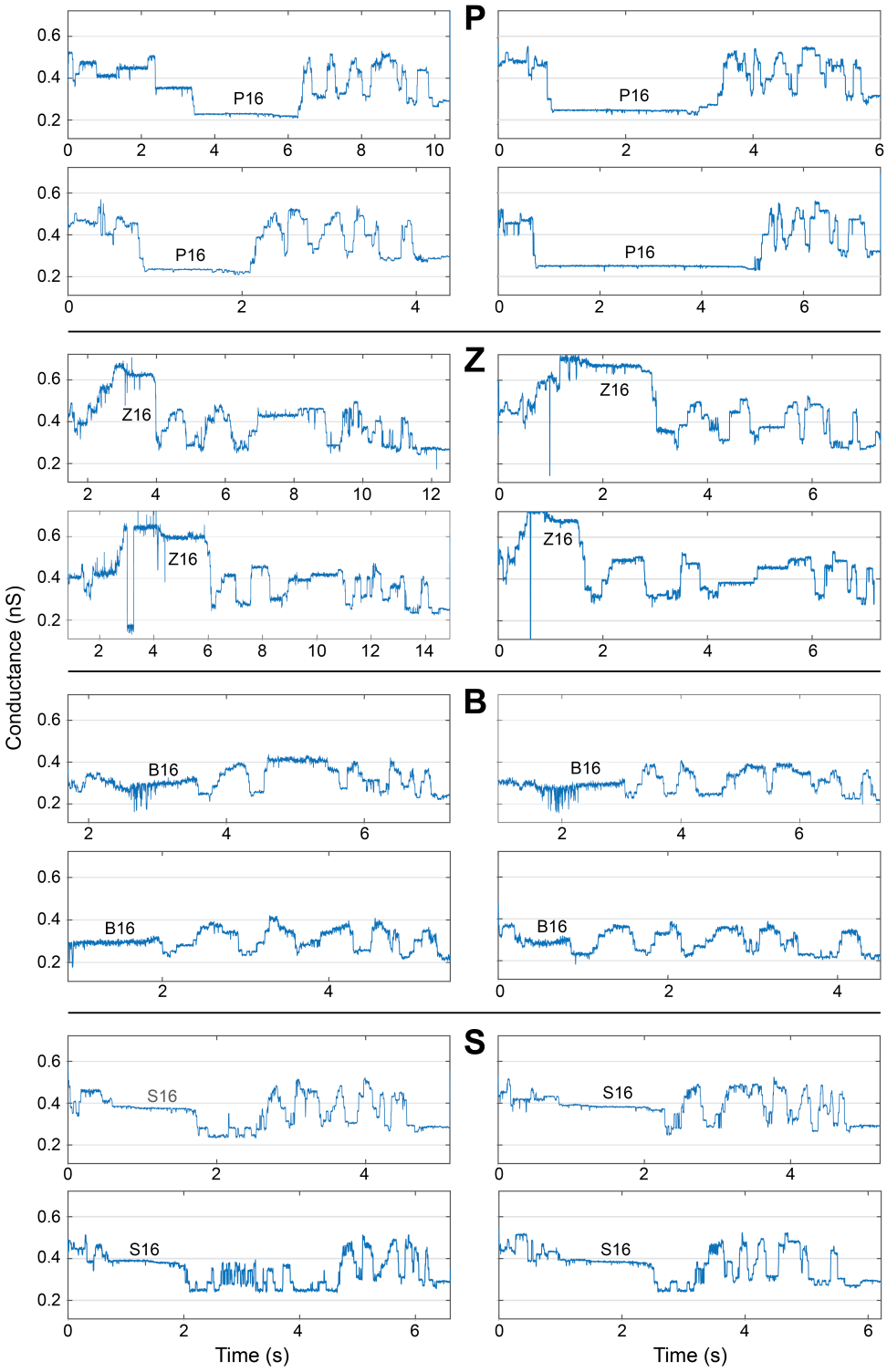

Figure B.1: Representative nanopore conductance traces for each non-standard base homopolymer tested. The 16 nucleotide homopolymer region is indicated by "N16" and is visible in the traces as a long semi-flat region of signal.

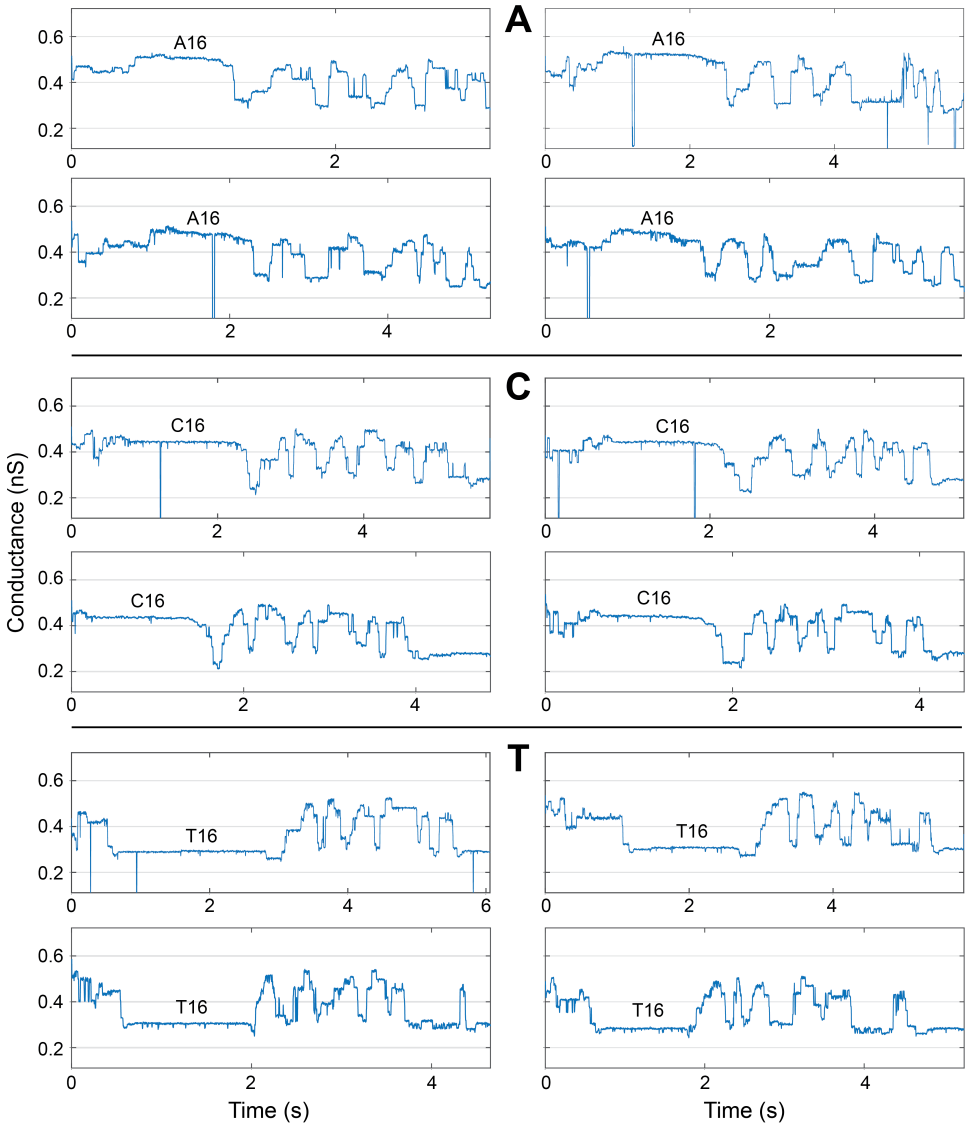

Figure B.2: Representative nanopore conductance traces for each standard base homopolymer tested. The 16 nucleotide homopolymer region is indicated by "N16" and is visible in the traces as a long semi-flat region of signal.

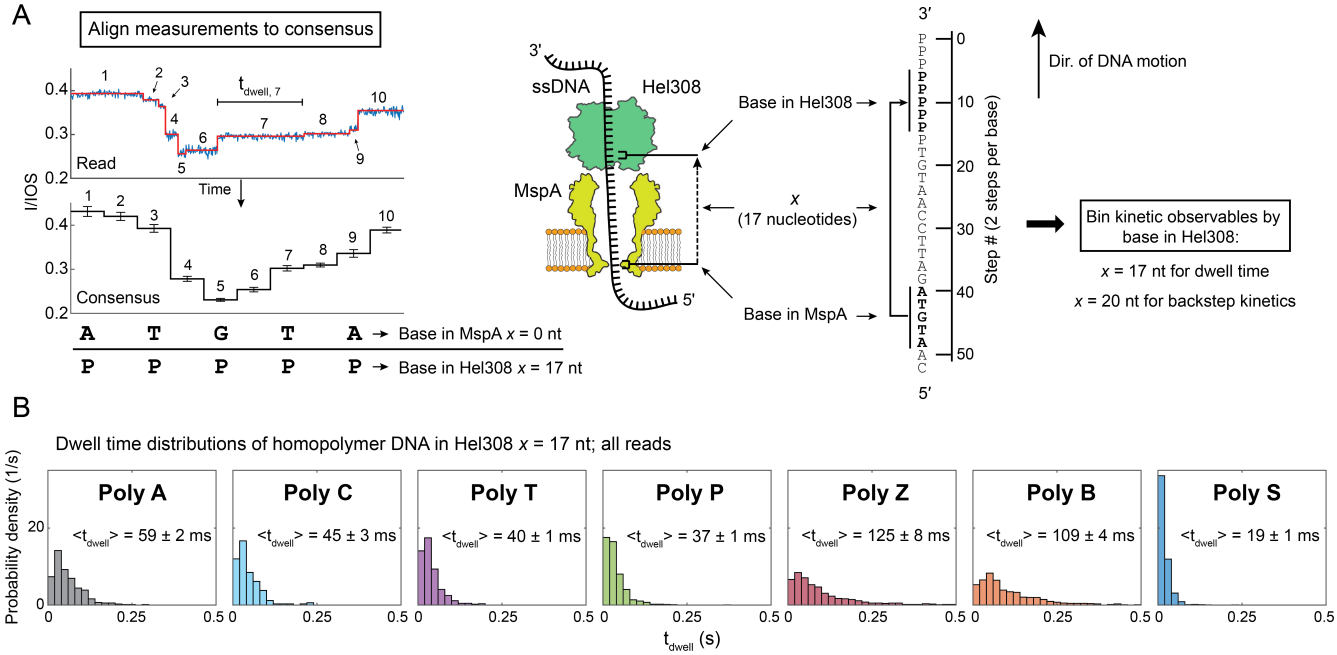

Figure B.3: Overview of kinetic analysis with nanopores (nanopore tweezers). (A) Segmented reads are aligned to a consensus and measurements of ion current, dwell time, and backwards steps are assigned to consensus positions. The kinetic measurements are binned by the bases within Hel308 (x = 17 nt and x = 20 nt upstream of the pore constriction for dwell time and backsteps respectively as in [55]. (B) Dwell time distributions of homopolymer DNA within Hel308 at x = 17 nt show dependence on the underlying sequence. Hel308 has a step size of 2-steps-per-nucleotide in our setup as established in [51]. Consensus errors are standard deviation; dwell time errors are SEM.

B.3 Characterizing Hel308 strand dissociation

We estimated the instantaneous probability of dissociation P_{dis} as a function of enzyme position by tracking the proportion of measured translocation events that dissociate during each observed enzymatic step along the DNA template (Equation B.1). Here x is the enzyme position along the DNA template in half-nt increments, $d(x)$ is the number of enzyme dissociations observed at position x, and $n(x)$ is the number of enzymes found to be still bound to the template at position x. We estimated the asymmetric uncertainty of P_{dis} using the Wilson Score Interval (Equation B.2). We also constructed Kaplan-Meier estimators [70] of the survival function $S(x)$ (Equation B.3) using the same parameters.

$$\hat{P}_{dis}(x) = \frac{d(x)}{n(x)} \tag{B.1}$$

$$\delta\hat{P}_{dis}(x) = \left[\hat{P}_{dis}(x) + \frac{1}{2n(x)} \pm \sqrt{\frac{\hat{P}_{dis}(x)(1-\hat{P}_{dis}(x))}{n(x)} + \frac{1}{4n(x)^2}}\right] \frac{1}{1+\frac{1}{n(x)}} \tag{B.2}$$

$$\hat{S}(x) = \prod_{i:x_i \leq x} (1 - \frac{d(i)}{n(i)}) \tag{B.3}$$

where x is the enzyme position along the DNA template in half-nt increments, $d(x)$ is the number of enzyme dissociations observed at position x, and $n(x)$ is the number of enzymes found to be still bound to the template at position x.

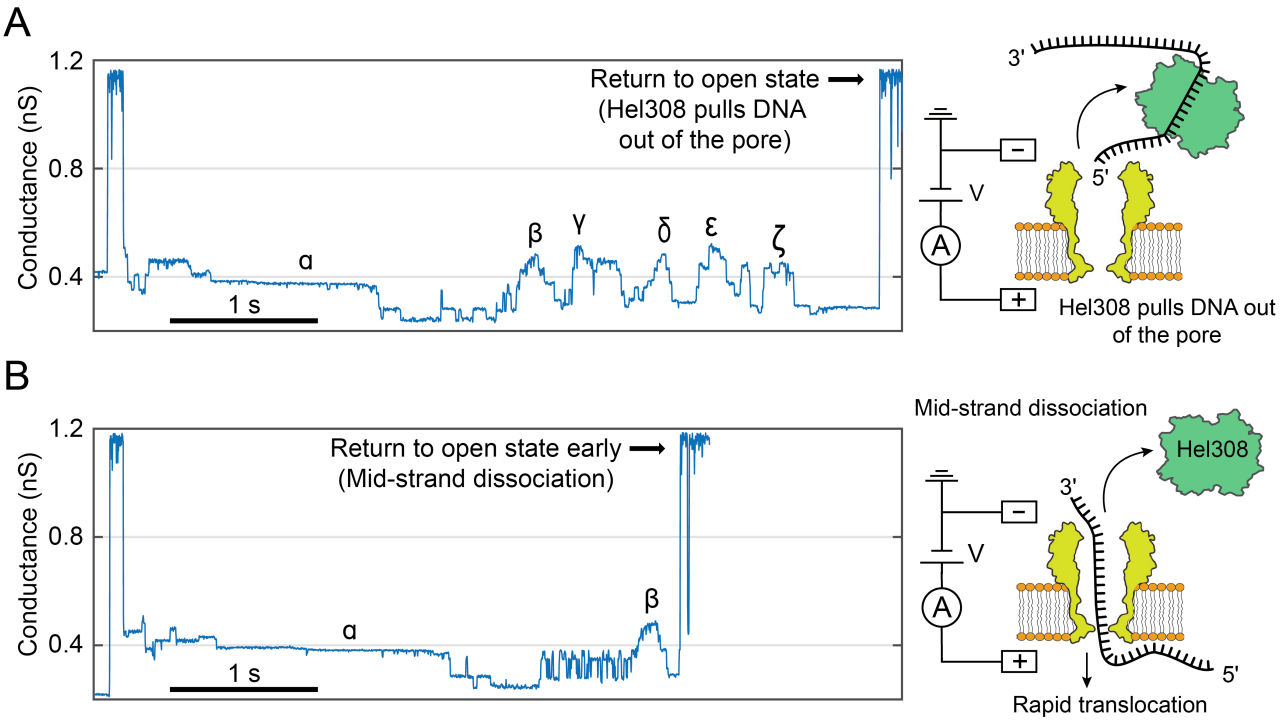

Figure B.4: Dissociation event example. (A) Full length conductance trace terminating with Hel308 pulling ssDNA out of the nanopore. (B) Partial length conductance trace due to putative mid-strand dissociation of Hel308 from the DNA template, followed by rapid translocation of the free ssDNA. We identify dissociation of Hel308 from a DNA template during a translocation event by observing the position on the consensus at which conductance trace returns to its open state value. Common conductance patterns within the traces are highlighted by Greek letters.

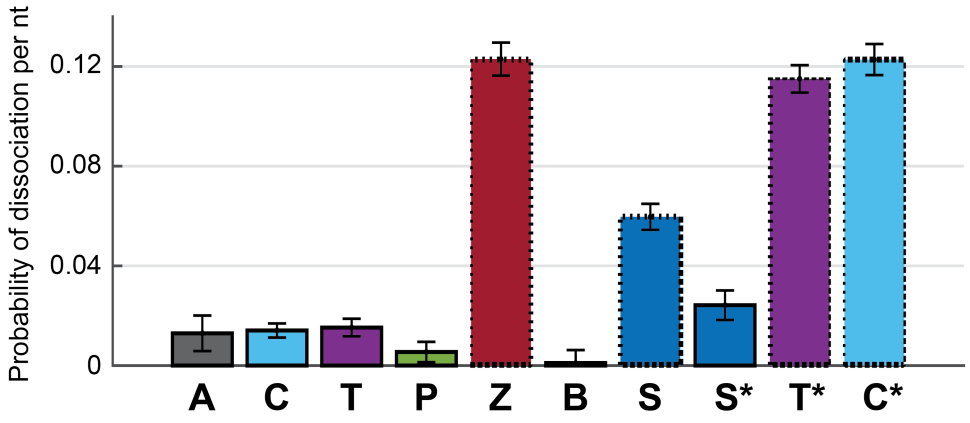

Figure B.5: Mean dissociation probability per nucleotide for each C-glycoside (dashed bar) or N-glycoside (solid bar) homopolymer with standard error of the mean. Base analogues are indicated by *.

B.4 LCMS analysis and quality control

We used liquid chromatography-mass spectrometry services from Novatia LLC to investigate our DNA sample purity. Most samples had a chromatogram dominated by one main elution peak corresponding to the molecular mass of the expected DNA strand. Poly B was the exception with two major peaks, which we believe is explained by the 16 consecutive **B** nucleotides forming quadruplex or pentaplex structures [67][68]. This is reinforced by the observation that the dominant molecular species in both elution peaks matched the molecular mass of the target Poly B strand.

The most abundant chemical adduct observed in the mass spectrums was a single cyanoethyl group attached to a strand, followed by single isobutyrl or single benzoyl adduct. Presumably these adducted strands made it through the PAGE purification process since a single adduct adds negligible mass to the 90 nt DNA substrate. We observed strands missing a single nucleotide or a terminal phosphate, but these would not have a significant effect on our dissociation results. We did observe a small quantity of high mass molecular species and interpreted them as unknown or multiple adducts.

The overall sample purity estimated using only the target molecule overestimates the molar proportion of strands that could bias our measurements of Hel308 dissociation, as impurities such as DNA fragments < 20 nt would not be bound by Hel308 and minor impurities lacking a single nucleotide or terminal phosphate would not cause early Hel308 dissociation in a way significant to our measurement. Larger DNA fragments (20 – 80mers) would result in a false dissociation signal, and we cannot rule out the effect of strands with large chemical adducts on Hel308

dissociation. Thus these elements must be taken into account when assessing the systematic effects of sample purity in our dissociation measurements.

To estimate the proportion of DNA strands that could significantly bias our dissociation measurement, we combined the proportion of all adduct strands (excluding adducts identified as originating from LCMS) with the proportion of all significantly truncated strands (20-80mers) from at least 90% of the total chromatogram area (Table B.3). We believe this provides the best estimate of the molar percentage of DNA sample impurities that could bias our measurements.

The impurity percentage estimated for all samples was $19 \pm 9\%$. This level of impurity is insufficient to explain the increased Hel308 dissociation observed with C-glycosides, in which 50-80% of reads ended prematurely. The observed presence of chemical adducts and DNA fragments from the solid-phase synthesis may however explain the discrepancy between Hel308 dissociaton measured on synthetic ACGT DNA (this study) and genomic DNA [52].

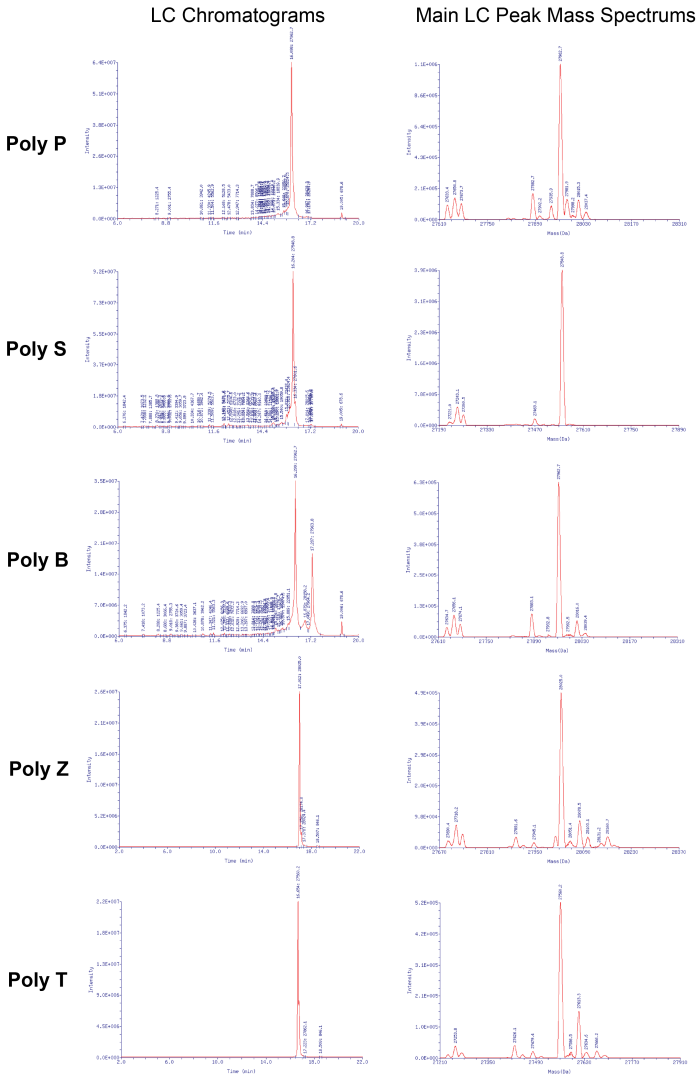

Figure B.6: Liquid chromatography chromatograms and mass spectrums of the largest-area elution peak for each of the tested samples. We interpret the double chromatogram peak observed in the Poly B sample to be due to quadruplex/pentaplex secondary structure formed by consecutive **B**'s. Common minor peaks observed in the mass spectrums of the main elution peaks in all samples include strands lacking a terminal phosphate or single nucleotide, A depurination, and chemical adducts from the solid-phase synthesis process.

Name	Impurity percentage
Poly T	28.2 ± 0.2
Poly P	10.9 ± 0.7
Poly Z	20.6 ± 0.4
Poly B	18.9 ± 1.6
Poly S	19.4 ± 1.8

Table B.3: Molar percentage of adducted and truncated strands that may affect our measurements of Hel308 dissociation. Includes the proportion of all adduct strands (excluding adducts identified as originating from LCMS) with the proportion of all significantly truncated strands (20-80mers) from at least 90% of the total chromatogram.

Appendix C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

C.1 Extended Materials and Methods

C.1.1 Pore Establishment

A single M2-NNN MspA nanopore was established in a 1,2-di-O-phytanyl-sn-glycero-3-phosphocholine (DOPHPC) lipid bilayer using methods that have been well established [48]. Lipids were from Avanti Polar Lipids.

C.1.2 Operating Conditions

All experiments were run at 400 mM *trans* [KCl] and 400 mM *cis* [KCl] with 10 mM HEPES at pH 8.0 and 10 mM [$MgCl_2$] (*cis* only) at temperature 37 ± 1 degrees Celsius. Once a single M2-NNN MspA nanopore was established, a buffer with the above conditions along with ATP was perfused to the *cis* well. ATP was ordered from Sigma Aldrich. The perfusion is done to maintain constant concentrations of the reactants/products in the reaction volume. DNA, DTT, and Hel308 were added to final concentrations of 10 nM, 1 mM, and 200 nM, respectively. Reactants/products were re-perfused every 45 minutes.

C.1.3 Proteins

M2-NNN MspA (accession number CAB56052.1) were prepared as described previously [48][51]. We used Hel308 from Thermococcus gammatolerans EJ3 (accession number WP_015858487.1), which expressed using standard techniques by

in-house facilities.

C.1.4 DNA preparation

AEGIS phosphoramidites (Firebird Biomolecular Sciences LLC. www.firebirdbio.com) were used in solid phase oligonucleotide synthesis on an ABI 394 instrument to make the non-standard DNA molecules. Oligonucleotides were purified on 10% denatured PAGE and then desalted on C18 Sep-Pak cartridge. DNA strands were suspended in 100 mM KCl at 20 μM concentration. Sequences were then mixed with their complement and were annealed at 90°C and then decreased step-wise by $\sim$ 10°C per minute to 4°C. Once annealed, these sequences were diluted to 1 μM for addition to the experiment. A complete list of DNA strands used in this work is given in Table C.1).

C.1.5 Chemical names

P 2-amino-8-(1′-β-D-2′-deoxyribofuranosyl)-imidazo-[1,2a]-1,3,5-triazin-[8H]-4-one

Z 6-amino-3-(14′-β-D-2′-deoxyribofuranosyl)-5-nitro-1H-pyridin-2-one

S 3-methyl-6-amino-5-(1′-β-D-2′-deoxyribofuranosyl)-pyrimidin-2-one

B isoguanine 6-amino-9[(1′-β-D-2′-deoxyribofuranosyl)-4-hydroxy-5-(hydroxymethyl)-oxolan-2-yl]-1H-purin-2-one

C.1.6 Data Acquisition

Data was acquired with custom labview software on an Axopatch 200B amplifier at 50 kHz, and downsampled by averaging to 5 kHz. We applied a 200 Hz, 100 mV peak-to-peak triangle waveform in addition to a constant 150 mV DC offset. Because

we filter the input voltage with a 3.5 kHz lowpass filter, the voltage waveform spans only $\sim$97 mV.

Name	Sequence (5′ - 3′)
Map seq 1	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSBZZZPPZPBPSZSBZSZPBSZSZBSZPGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 2	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSZPSPBZZBBZPZZPBBBSSZBPSPPBSGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 3	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSBSPBBPSBPPSZPZSPBPPBBZSPZZZGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 4	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSBZZZPPZPBPSZSBZSZPBSZSZBSZPGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 5	/phos/TTTTTTTTTTCCTTCCTTAGATTCCZZZBSPPPZSBBZZPZBPZZSPSSBZPBGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 6	/phos/TTTTTTTTTTCCTTCCTTAGATTCCZPBZBZPPBZSSBSZBZZSZSPPSPSZZGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 7	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSZZSSSPSPZBSSSBBPPZBZBSBSSPBGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 8	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSPBSSBPSSSZPPPBPZPZPSZBBSPSBGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 9	/phos/TTTTTTTTTTCCTTCCTTAGATTCCPSBZBPPPSSZZZSBSBPZSZZPSBBBPGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 10	/phos/TTTTTTTTTTCCTTCCTTAGATTCCBBPZBBBZBBPBBSBBSZZBZSBPBPBSGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 11	/phos/TTTTTTTTTTCCTTCCTTAGATTCCSSZBZSSBPBBBSSZPSPBSBZZBPSPPGAGAGAGAGAGAGAGAGAGAGGAAGCATTCTCCTTAT
Map seq 12	/phos/TTTTTTTTTTCCTTCCTTAGATTCCAATGTAGATTCCAATGTBBBBPPPPSSSSZZZZCACGGCAGACAACTCATGCAGGAAGCATTCTC
Test seq 1	/phos/TTTTTTTTTTCCTTCCTTAGATTCGZSBBSZPPBSBBSPSSBSZBPBZPSZPZGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 2	/phos/TTTTTTTTTTCCTTCCTTAGATTCGZSBSPZPPZBPZZBSBBBZSSSPZBZZZGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 3	/phos/TTTTTTTTTTCCTTCCTTAGATTCGBBBZBSSBZBPZSSPZSBZBZPBPBBPBGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 4	/phos/TTTTTTTTTTCCTTCCTTAGATTCGPSPZSSZZZBZPSBPSPSZPSSSPSBSSGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 5	/phos/TTTTTTTTTTCCTTCCTTAGATTCGPZSBZSZSBPZSZBSZSPBPZPPZZZBSGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 6	/phos/TTTTTTTTTTCCTTCCTTAGATTCGSBPPSPBZSSBBBPZZPZBBZZBPSPZSGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 7	/phos/TTTTTTTTTTCCTTCCTTAGATTCGSBSPPZPSSPZBPBZZPPSSZSSZPSBPGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Test seq 8	/phos/TTTTTTTTTTCCTTCCTTAGATTCGBPZBSSPSZZPBZPZZBBSZPPPSBZZZGAGAGACCTCCTCATCCTCCCGAAGCATTCTCCTTAT
Map seq complement	TGCTTCCTCTCTCTCTCTCTCTCTCTTTTTTTTTTTTTTTTTTTT/chol/
Test seq complement	TGCTTCGGGAGGATGAGGAGGTCTCTCTTTTTTTTTTTTTTTTTTTT/chol/
Special complement	CCTGCATGAGTTGTCTGCCGTGXXXX/chol/

Table C.1: Complete list of DNA sequences used for sequencing experiments. /phos/ = 5′ phosphate, /chol/ = 3′ cholesterol tag, X = 18-carbon spacer. Map seq complement was annealed to all Map seq strands except Map seq 12. Test seq complement was annealed to all Test seq strands. Special complement was annealed to Map seq 12.

Name	Number of reads
Map seq 1	22
Map seq 2	29
Map seq 3	35
Map seq 4	37
Map seq 5	34
Map seq 6	41
Map seq 7	32
Map seq 8	23
Map seq 9	40
Map seq 10	47
Map seq 11	48
Map seq 12	37
Test seq 1	15
Test seq 2	17
Test seq 3	12
Test seq 4	18
Test seq 5	29
Test seq 6	23
Test seq 7	14
Test seq 8	20

Table C.2: Summary of all nanopore reads collected for each DNA template used in Chapter 4.

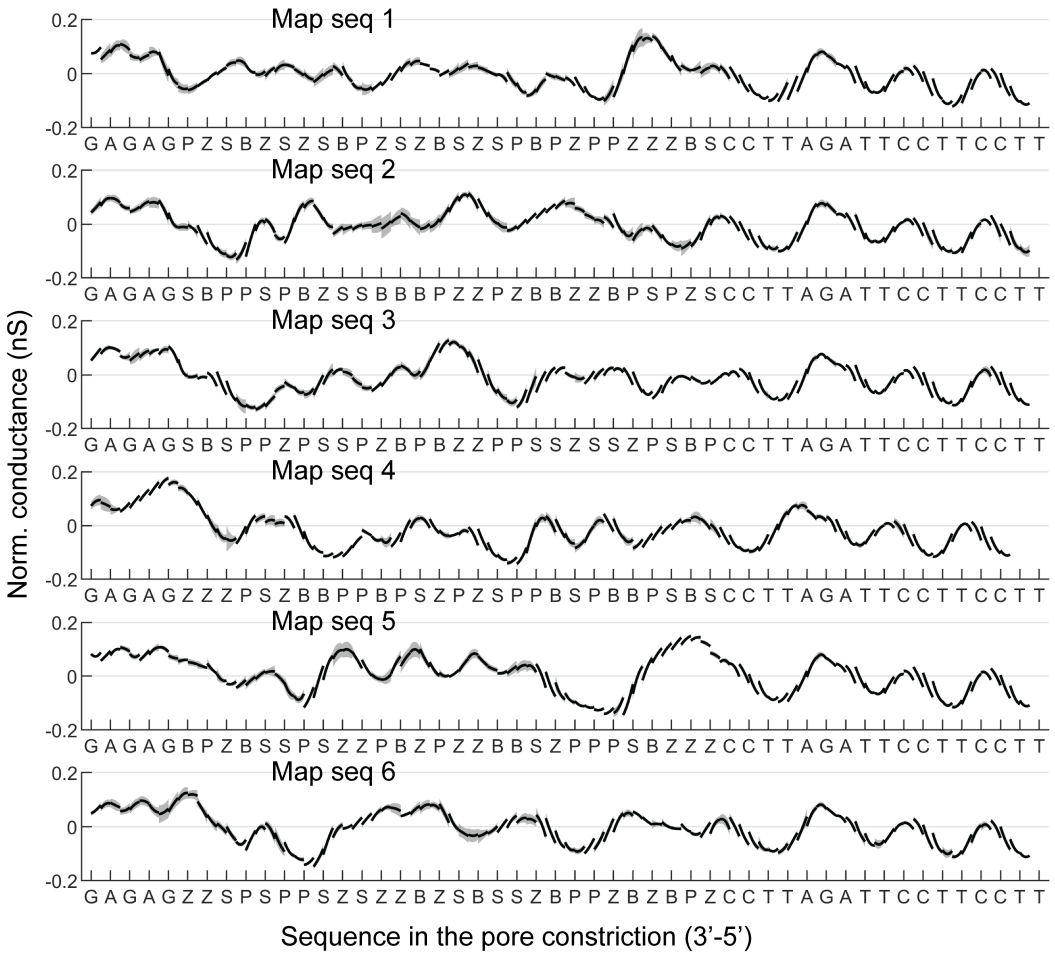

Figure C.1: Sequence-aligned consensus traces for map building strands #1-6. Shaded regions indicate standard deviation of the consensus state.

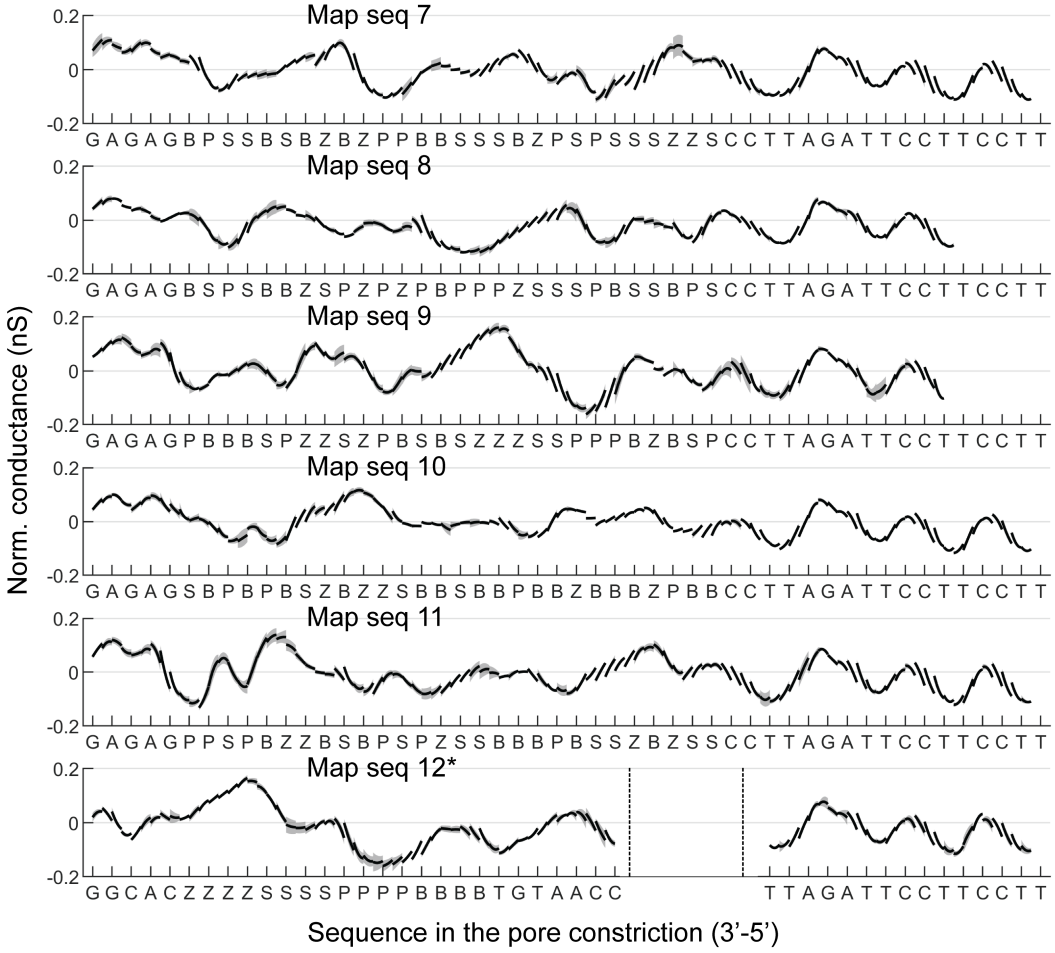

Figure C.2: Sequence-aligned consensus traces for map building strands #7-12. The consensus trace for Map seq 12 spans two separate parts of the sequence. Shaded regions indicate standard deviation of the consensus state.

C.2 K-mer model covariance estimation

Shrinkage approach for ALIEN 3-mer model.

$$\Sigma_i^{global}(\lambda) = (1-\lambda)\Sigma_i + \lambda\Sigma_{global} \tag{C.1}$$

where $\Sigma_i^{global}(\lambda)$ is the blended covariance matrix of k-mer i, Σ_i is the covariance matrix of k-mer i, Σ_{global} is the estimated global covariance matrix of all measurements, and λ is the blending parameter.

$$\Sigma_{global} = \frac{1}{N-k}\sum_{i=1}^{k}(N_i - 1)\Sigma_i \tag{C.2}$$

where N is the total number of measurements, k is the number of k-mers in the model, and N_i is the number of measurements for k-mer i.

$$\lambda = \begin{cases} \frac{5-N_i}{5}, & \text{if } N_i < 6 \\ 0, & \text{otherwise} \end{cases} \tag{C.3}$$

Global covariance estimation for ALIEN 4-mer model.

$$\boldsymbol{g}_{context\,1} = (\boldsymbol{g}^1_{context\,1},\ \boldsymbol{g}^2_{context\,1},\ \cdots,\ \boldsymbol{g}^N_{context\,1}) \tag{C.4}$$

$$\boldsymbol{g}_{context\,2} = (\boldsymbol{g}^1_{context\,2},\ \boldsymbol{g}^2_{context\,2},\ \cdots,\ \boldsymbol{g}^N_{context\,2}) \tag{C.5}$$

$$\Sigma_{global} = cov(\boldsymbol{g}_{context\,1} - \boldsymbol{g}_{context\,2}) \tag{C.6}$$

where N is the number of duplicate k-mer measurements in separate sequence contexts.

C.3 Evaluating random basecalling

To evaluate the performance of our k-mer model basecalling, it would be insightful to compare k-mer model accuracy to accuracy achieved with purely random basecalls. However, the expected accuracy of purely random basecalling is not trivial when using local sequence alignments that allow for insertions and deletions. Since local alignments do not penalize gaps at the beginning or ends of the reference sequence, shorter random base read lengths will have intrinsically higher accuracy than longer read random base lengths.

To give the best possible comparison between k-mer basecalling and random, for each read basecall we generated a distribution of 50 random ALIEN sequences of identical length to the read basecall. We then aligned these random reads to the reference sequence to calculate accuracy. Generating a distribution of random sequences for every individual read has the advantage of averaging the random accuracy, while the identical lengths of the random and read basecalls accounts for the read length dependent effect on accuracy.

Appendix D

A LEXICOGRAPHIC STUDY OF GENETIC ALPHABETS

As it turns out, the letters used to symbolize nucleic acid bases are a subset of the same letters used to construct English words. Here I attempt to study the lexicographic properties of the DNA alphabets I have thus far encountered, namely ACGT (4-letter), ACGT**PZBS**, (8-letter *hachimoji*), and ACGT**PZBSKXJV** (12-letter *junimoji*). As a reference for English words, I opted to use the dictionary of valid words from popular word game *Scrabble*™.

I start by counting the number of valid words capable of being formed by the letters in these three genetic alphabets (Figure D.1 Top). Unsurprisingly, the number of words formed by the three alphabets grows with the number of letters each alphabet has. Notably, the letter **S** has a disproportiately large effect on the number of words formed as it introduces the plural forms of most words.

Because *Scrabble*™ involves scoring points based on letter point values summed from letter contained within played words, I also investigated the total quantity of points able to be scored based on words formed from each alphabet (Figure D.1 Bottom, Table D.1). While the letter **S** expands the number of words, it has a low letter point score (1). In contrast, the extra letters **KXJV** in the 12-letter alphabet have a high point score which, in addition to the longer words the 12-letter alphabet is capable of producing, contribute to the 12-letter alphabet having a larger total point sum than might otherwise be expected.

Of particular note is that all alphabets contain only a single vowel (A), severely limiting the word count for all genetic alphabets. This insight highlights the need for

chemists to develop new synthetic bases that can be notated with other vowels such as E or I.

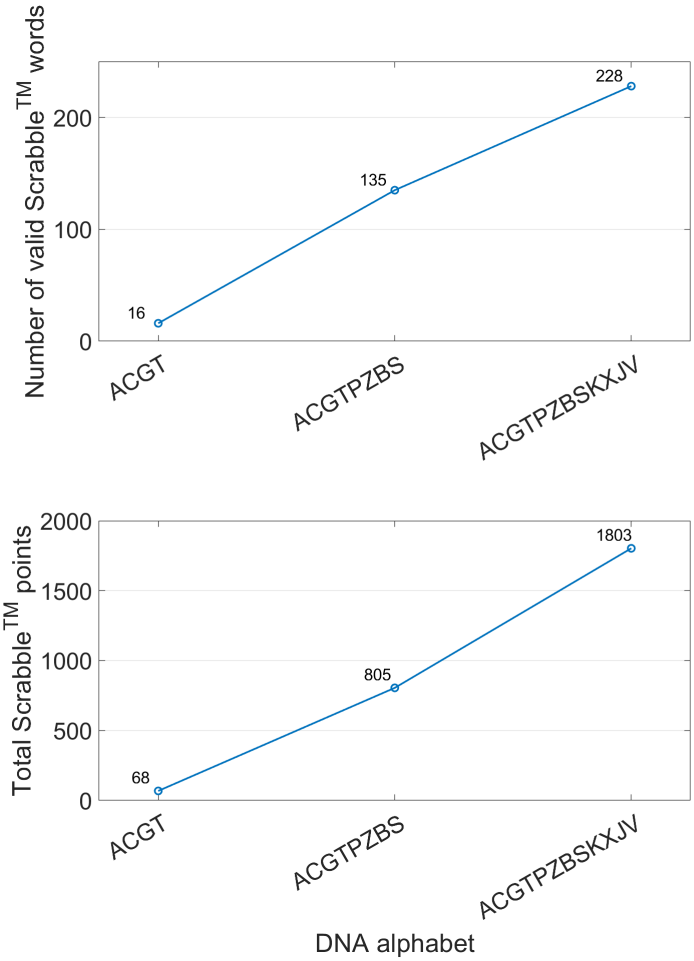

Figure D.1: (Top) Number of valid English words able to be formed by the letters of various genetic alphabets. Validity in this context refers to words allowed in the popular word-building game *Scrabble*™. The letter **S** has a disproportionately large effect on the number of possible words for alphabets containing it, as it enables the plural forms of most words. (Bottom) The total points as scored in *Scrabble*™ summed from every valid word in each alphabet. The *Scrabble*™ dictionary was sourced from https://scrabble.merriam.com/.

Letter	Point value
A,S,T	1
G	2
B,C,P	3
V	4
K	5
J,X	8
Z	10

Table D.1: Letter point values used by the boardgame *Scrabble*™ to score words. We only show letters used in the genetic alphabets studied here.