UNIVERSITY OF CALIFORNIA
SANTA CRUZ

# METHODS FOR ANALYSIS OF NANOPORE DNA SEQUENCING DATA

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

DOCTOR OF PHILOSOPHY

in

CHEMISTRY

by

**Arthur C. Rand**

June 2017

The Dissertation of Arthur C. Rand
is approved:

---

Michael D. Stone, Chair

---

Mark Akeson

---

Seth M. Rubin

---

Benedict Paten, Ph.D

---

Tyrus Miller
Vice Provost and Dean of Graduate Studies

# Table of Contents

# List of Figures

# List of Tables

**Abstract**

Methods for Analysis of Nanopore DNA Sequencing Data

by

Arthur C. Rand

Technology guides the practice of scientific inquiry. In the biological sciences, DNA sequencing has encouraged the commingling of traditional experimental biology and computer science. In this thesis, I describe computational and biochemical methods for DNA sequencing technology. Nanopore DNA sequencing is an single-molecule technology that shows promise in the area of read lengths, instrument portability, and, as shown in this work, chemical modification detection. A nanopore sequencing device contains a nanometer-sized pore that separates two electrolyte buffer reservoirs. A voltage potential is applied across the nanopore and the device records the ionic current through pore. As DNA polymers translocate through the pore they modulate the ionic current by partially obstructing the pore in a sequence-dependent manner. In Chapter 2 I describe a hidden Markov model for the nanopore ionic current and how a hierarchical Dirichlet process can be used to model new non-canonical (modified) bases affording a HMM-HDP model. In chapter 3 I show how the model can detect multiple chemical modifications to DNA. The last section of the paper describes a biochemical method to re-read DNA sequences using a nanopore and a helicase enzyme.

To my wife, Jennifer.

## Acknowledgments

This work is the result of me "doing" a Ph.D., but is really an aggregate of lots of experiences shaped by my lab mates and mentors during my time as a graduate student. I was fortunate to have talented lab mates during my career, especially Andrew Smith and Miten Jain, with whom I enjoyed lots of productive and stimulating conversations. I started my career as a "wet-lab" biochemist, and finished it as a "computational biologist" a transformation that I did not forecast. For this, I need to thank Professor Kevin Karplus and the BME205 class that he famously taught for changing the course of my career during and after graduate school. However, the person with the largest influence on my work was Dr. Benedict Paten. I approached Benedict as a third year student looking for a "little programming project" and was handed the project that ultimately became the cornerstone of my thesis. Benedict constantly challenged me and I am incredibly grateful for his tutelage. This work benefited greatly from the influence of my advisor Professor Mark Akeson whose guidance allowed me to explore new scientific endeavors while constantly pushing me to do high quality work.

# Chapter 1

# Introduction

Sequencing DNA allows for precise biological investigation. The datasets produced by modern DNA sequencing experiments are a rich source of information and often out live their original study. This work focuses on nanopore-based DNA sequencing. Nanopore sequencing belongs to a class of technologies often referred to as *single-molecule* or *third-generation* technologies. The practical benefits of these technologies are long read lengths, device portability, speed, and DNA modification detection. The last of which is the focus of this thesis.

## 1.1 Single-molecule investigation of nucleic acids

Sequencing DNA is fundamentally about analysis of organic molecules. Limiting the search to nucleotide polymers only serves to narrow the search space. *Single-molecule* analysis of biomolecules, however, is somewhat unique compared to other analysis methods where an ensemble of molecules are observed together. Molecular

tweezers have allowed researchers to probe the mechanical characteristics of individual DNA molecules [34] and DNA processing enzymes [32, 27, 38, 14]. Tweezers allow for direct observation of a biological complex at angstrom scale in real time (on the order of milliseconds) as well as record force measurements on the order of pico Newtons [39]. Another, in many ways, unique, single molecule technique used in DNA sequencing uses zero-mode waveguide wells to restrict the observable space to a volume where individual DNA polymerases can be monitored [28].

Traditionally, nanopore sensors were used to answer similar questions to these methods. The basic setup of a nanopore sensor can be seen in Figure 1.1. An artificial membrane partitions two wells containing electrolyte buffer. A single nanopore, usually an engineered protein, is isolated in the membrane and allows for the passage of certain biomolecules and electrolytes. A voltage is applied across the membrane that induces an ionic current through the nanopore.

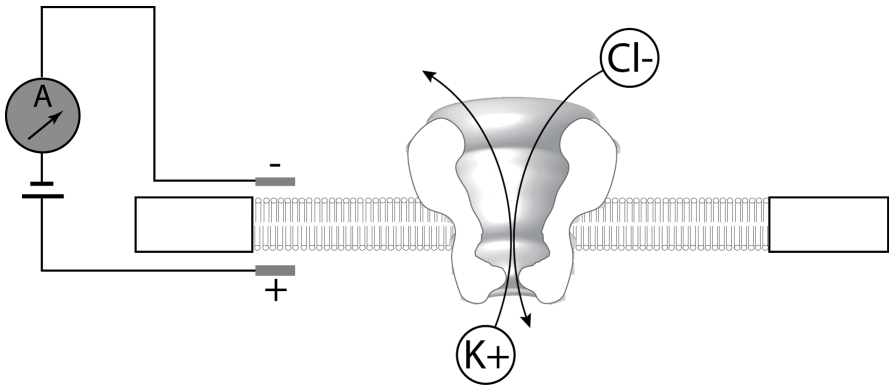


Figure 1.1: Traditional single-molecule nanopore setup showing a protein pore embedded in an artificial membrane with an ionic current flowing through the nanopore.

In the case of investigations of nucleic acids, nucleic acid polymers (hereafter

referred to as a *substrate*) are added to one partition of the apparatus and electrophoretically driven through the nanopore. The ionic current (on the order of pico-amperes) is monitored in real time using a patch clamp amplifier [1]. When a substrate passes through the nanopore from the *cis* well to the *trans* well, it partially obstructs the passage of the ions though the nanopore sequence-specific way. These data are collected as a time series and referred to as a *trace*. These traces can be decoded into the underlying nucleotide sequence.

## 1.2 Nanopore characterization of DNA polymers

A simple setup where nucleic acid polymer translocation is driven by the voltage alone results in translocation times on the order of tens of microseconds per nucleotide. These traces contain enough information to distinguish between molecules of differing lengths, but not to distinguish nucleotide sequence beyond homopolymer composition [1]. To gain more information about the sequence of the substrate, it is necessary to slow the translocation of the substrate through the nanopore. The most successful method to date has been to couple the nanopore sensor with a DNA (or RNA) metabolizing enzyme. This led many researchers to search for the best enzyme/nanopore combination [20, 21, 10, 19]. A side-effect of these studies was the increased knowledge of the exact biophysics of these enzymes at high precision [12, 13]. A highly desirable system would have the following characteristics:

- The nanopore must have a shape such that different nucleotide sequences produce

distinguishable changes in the ionic current.

- High *processivity* of the motor enzyme, meaning that the entire motor/DNA/pore complex metabolizes a large amount of the substrate before the complex dissociates.

- Reproducible behavior of the system.

In 2012, a collaboration between the nanopore groups at University of California at Santa Cruz and University of Washington yielded two papers showing that the DNA polymerase from bacteriophage phi29 was especially adept to this task. These papers showed that the polymerase was able to consistently slow the translocation of the DNA to 2.5-40 nucleotides per second, slow enough to resolve individual "events" or "levels" due to single nucleotide steps of the substrate through the pore [36, 9]. These papers also highlighted the opportunistic use of creative DNA architectures to allow for the enzyme to process the DNA in a favorable way. In these papers, for example, the main challenge was pr prevent the DNA polymerase from acting *in bulk* before the ternary complex is captured by the pore. This is especially important because the DNA polymerase chemically modifies the substrate, and thus cannot be re-analyzed. The solution in this case was to use the *blocking oligo* a hybridized partially complementary oligonucleotide that prevented the DNA polymerase from replicating the DNA strand but did not prevent it from binding to the substrate. The blocking oligonucleotide is removed when the DNA/enzyme complex is captured by the nanopore and is allowed to process.

Subsequent papers would go on to show that the nanopore MspA (introduced by the UW group) was sensitive enough to distinguish between modified and unmodified cytosine bases within a known DNA context [52, 46, 25]. The UW group also showed that the phi29 DNA polymerase is capable of metabolizing strands of almost a kilobase in length and that the traces themselves could be decoded (*base called*) into DNA sequence reads [26].

## 1.3 Introduction of the Oxford Nanopore Technologies MinION

The 2014 paper by Laszlo *et al.* was the most advanced research-grade demonstration of the potential of nanopore sensors as a relevant DNA sequencing technology. However, later that year Oxford Nanopore Technologies (ONT) would soft release their commercial nanopore-based sequencing instrument, the MinION, to approximately one hundred research labs. The MinION uses a application specific integrated circuit containing an array of individually controlled nanopore sensors that each record ionic current traces. ONT licensed patents from the University of California regarding what has become known as *stand sequencing*. Strand sequencing is the specific approach originally described by the UCSC nanopore group where the DNA strand is fed through the nanopore sensor controlled by an enzyme. Alternative approaches using enzymes to destruct (or construct) the DNA polymer have also been proposed [18, 3]. The instrument itself is the size of a candy bar and is powered by a USB 3.0 port. It streams

the ionic current traces to a laptop where, at least initially, the traces are uploaded to ONT's proprietary cloud-based service, Metrichor, quality checked, base called, and are then downloaded back to the laptop.

A DNA sample, usually referred to as a *library* is prepared by first shearing the genomic DNA into fragments followed by repair and ligating sequencing adapters on either end (see [22] figure 1a.). These adapters allow for the motor enzyme (a helicase in the case of the MinION) to mediate the DNA translocation through the nanopore. Initially, the strand sequencing accuracy was not very good, with a mean base identity of 75%, to improve the accuracy for a single molecule a hairpin adaptor is ligated to the substrate effectively making the molecule into one long DNA strand. This allows the motor enzyme to process both strands of the DNA duplex into what are called $2D$ reads. The increased accuracy comes from an alignment of the two individual strands [22]. The effective *rereading* of the strand to increase accuracy was part of the motivation to develop specific methods to reread a single molecule of DNA (discussed in Chapter 4).

Data produced by MinION has a few notable features. First, similar to SMRT sequencing, the DNA does not require PCR amplification prior to sequencing. By sequencing DNA directly from the cell, chemical modifications such as methylation are retained on the substrate. Second, the raw ionic current data is packaged along with the reads. Although it is not uncommon for sequencing instruments to provide the raw measurements along with the processed sequencing data, additional *post hoc* analysis of the MinION ionic current traces has been a rapid area of research and is the subject of

Chapter 2.

During the first few years of the early access program dozens of papers have been published touting the MinION's unique characteristics and performance. Many of these papers have leveraged the long reads produced by the instrument (on the order of tens of kilobases in length being typical at the time of this writing), portability, and speed. [41, 2]. The technology evolves rapidly, however, which while good news for researchers because the performance of the instrument is consistently improving, also means that researchers may rush to publish data and results before they are outdated. The first widely released edition of the sequencing workflow was termed $R7.3$. This version widely used the $2D$ sequencing technique and processed the DNA at 120 base pairs per second. At its apogee, sequencing reads had 85% base identity and a generally robust workflow. The base calling was done using a hidden Markov model (HMM) using the Metrichor cloud service. The initial DNA modification modeling (Section 3.5) was done on the $R7.3$ chemistry.

In early 2016, ONT released the $R9$ sequencing kits and associated workflows. The new version used a different pore, (CsgG from *E. coli*), and an increased sequencing speed of 240 base pairs per second. The base calling (still done in the cloud) was now done by a long-short term memory recurrent neural network (LSTM-RNN) a more advanced type of model which have been shown to be especially adept at machine translation tasks. These two changes improved the sequencing performance to consistently in the mid-90% base identity. The newer results showing modeling of DNA modifications in bacteria (Section 3.6 and 3.7) and human genomic DNA (Section 3.8) use the $R9$

chemistry.

## 1.4 Nanopore sequence analysis with hidden Markov models

Hidden Markov models (HMMs) are graphical models commonly used in bioinformatics due to their extensibility and probabilistic interpretation. Classic examples include protein structure prediction and sequence searching [24, 16]. These models can also be used for alignment, although they are somewhat less performant than the more commonly used Smith-Waterman based alignment algorithms [29].

A key advantage of using HMMs for biological sequence alignment comes from their probabilistic output, specifically the ability to derive a *posterior probability*. An abridged version of the description by [15] follows. The HMM is a form of a Bayesian network that represents a distribution over sequences of observations. Each node in the graph of the graph represents an unobservable (hidden) state of the system and the edges of the graph represent the probability of moving from one state to another. More formally, consider a sequence of observations $\mathbf{x} = \{x_i, \ldots, x_n\}$ and a set of symbols $b \in \mathbf{B}$. The hidden states are characterized by the *emission* probability table $\mathcal{E}$ which if given by: $\mathcal{E}(b) = P(x_i = b|\pi_j = k)$, where $\pi_j$ is the state at time $j$ in the path $\pi$ through the HMM lattice, $k$ is the identity of the hidden state (node) in the graph, and $i$, $j$ is the time step of the observations and the paths, respectively. The edges of the graph represent the *transition* probabilities of the system, which can be intuitively

thought of as the probability of moving from one hidden state to another. The transition probability table is given by $\mathcal{A}_{k \to l} = P(\pi_i = l | \pi_{i-1} = k)$. As described in Chapter 2 both of these probability tables can be *learned* from an input dataset.

The real power of the HMM is comes from utilizing these two probability tables to compute the full probability of a sequence of observations given the model affording the *forward/backward* HMM briefly described here. To calculate the full probability of the observation sequence, $P(\mathbf{x})$, we sum over all paths for the joint probabilities of the path and $\mathbf{x}$,

$$P(\mathbf{x}) = \sum_{\pi} P(\mathbf{x}, \pi) \tag{1.1}$$

A naive solution would run with time complexity $O(L \cdot n^L)$ (where $L$ is `x.length` and $n$ is the order of the HMM graph). A belief propagation method (implemented here) uses the recursion.

$$c_l(i) = \mathcal{E}_l(x_i) \sum_k c_k(i + o) \cdot \mathcal{A}_{k \to l} \tag{1.2}$$

Where $l$ is the current state, $k$ is a previous state, and $o$ is the time offset ($o \in \{1, -1\}$). The well known forward and backward variables are then $f_l(i) = c_l(i), o = 1$ and $b_l(i) = c_l(i), o = -1$. Calculating the forward and backward matrices of probabilities is performed using dynamic programming. From here we can calculate the posterior probability that a observation, $x_i$, was emitted from state $e_k$ by $P(\pi_i = k_j | x) = \frac{f_l(i) b_l(i)}{P(\mathbf{x})}$. This allows us to probabilistically infer latent features about the data.

## 1.5 Description of this work

In this work I develop new methods to improve the utilization of a developing third generation sequencing technology, nanopore sequencing. Nanopore sequencing is unique in the data stream it produces and is amenable to machine learning methods. Specifically, by implementing with multiple hidden Markov model (HMM) topologies and modeling paradigms (Chapter 2, Section 2.1) I show that a traditional aligning model can be used to probabilistically assign continuous ionic current segments (*events*) to a reference sequence. In this chapter at I describe a traditional pair-HMM, a HMM where the state transitions are conditionally dependent on the emissions of the previous state, and a one where the time duration of the event stream is considered. The goal of the modeling endeavour was not to simply assign events to the reference sequence but to use the aligned events to classify and detect various unknown qualities of the reference sequence, specifically base modifications. In Section 2.2 I show how the simple pair HMM can be augmented with a variable-order meta structure that allows for multiple different bases to be simultaneously aligned to any number of positions in the reference sequence. For variant detection (where we suspect that a base in the true sequence is different from that in the reference) I showed how this model can be used to sample multiple candidate sequences in a principled way to propose variants (Section 2.6). A major advantage of single-molecule sequencing, and nanopore sequencing, is that the substrate DNA does not require amplification prior to sequencing, meaning that the substrate retains any biological markers present in the cell. In Sections 2.2.2 through

2.4 I show how the emissions probabilities of the variable order HMM can be trained to learn new bases using a hierarchical Dirichlet process (HDP) in an iterative expectation maximization and Gibbs sampling procedure. Lastly, in Section 2.7 I show how this model can be scaled to compute on large datasets such as those produced when sequencing the human genome.

In Chapter 3 I show how the above model can be used to detect two cytosine methylation products (5-methyl cytosine, and 5-hydroxymethyl cytosine) and one adenine methylation product (6-methyl adenine). To date, few sequencing technologies have been able to directly detect and classify modified bases in genomic DNA. Currently, this includes standard ONT MinION protocols which have been designed to exclusively call canonical bases. I describe how aligned events can be classified with neural networks without using the variant calling power of the variable order HMM in Section 3.4. In Section 3.5 I describe the classification performance of the variable order HMM on synthetic oligonucleotides bearing one of the three cytosine variants. There are two methylation marks of importance in bacteria, 5-methyl cytosine in C**C**(A/T)GG contexts and 6-methyl adenine in G**A**TC contexts. In Section 3.6 I describe the performance of the model in detecting these two marks in a controlled plasmid system and *E. coli* genomic DNA. In Section 3.7 I show how the model is sensitive enough to detect dynamic changes in genome-wide methylation levels. Lastly, in Section, 3.8 I demonstrate the classification performance of the model on human genomic DNA at scale.

Chapter 4 contains work from early in my graduate career, specifically de-

velopment of biochemical techniques to use in nanopore sequencers. With the rapid development of the MinION and my shift in focus to computational biology this work become less relevant, however it was still the subject of multiple talks and presentations. In this chapter is describe how DNA can be reread using helicase enzyme coupled with a translocation preventing structure called a G-quadruplex. The G-quadruplex folds single stranded DNA into a knot that prevents the helicase from translocating in bulk, a technique similar to the "blocking oligo" originally used by Cherf *et al.* [9], but for a different enzyme. This work resulted in a patent for the adapter technique.

# Chapter 2

# Modeling of MinION signal data

Modeling of nanopore data has historically been a niche area of bioinformatic research [45, 51], but the commercialization of the technology in the form of the Oxford Nanopore Technologies (ONT) MinION has broadened the field to more researchers. Many early algorithms focused on *de novo* genome assembly and consensus building [33] hidden Markov models (HMMs).

The standard MinION DNA preparation protocol involves ligating a hairpin to one end of the DNA duplex so that both the template and complement strands are sequenced [22]. During sequencing, the MinION continuously records ionic current and then divides it into segments referred to as *events*. Both the ONT software and the method described here method model each event as a nucleotide string of length k, a k-mer. The length of the k-mer depends on the model used, which varies between sequencing protocols (k=6 for R7.3 and k=5 for R9). Each k-mer is associated with a distribution of ionic currents in picoamps (pA).

### Publication Note

The model described in this chapter is published here: [42] and the words herein are attributed to myself and my co-authors. Section 2.6 contains unpublished algorithm descriptions.

### Implementation Note

Code containing the implementation of these models and the cloud-based workflow described in Section 2.7 can be found at `https://github.com/ArtRand/signalAlign`, `https://github.com/ArtRand/toil-marginAlign`, and `https://github.com/ArtRand/toil-signalAli`

## 2.1 Pair hidden Markov model

During strand sequencing multiple nucleotides occupy the nanopore at one time and thus multiple nucleotides contribute to the ionic current. The motor enzyme moves the DNA one nucleotide at a time through the nanopore so $k-1$ of the nucleotides remain and one changes, resulting in a new k-mer in the nanopore. Following is the description of three models for aligning ionic current events (event sequence) to nucleotides (reference sequence such as a genome). The first model is adaptation of the traditional pair aligning HMM for use with continuous data types, the second is a reimplementation of the model by Loman *et al.* implemented as a forward/backward HMM, and the third is a generalized HMM that incorporates event duration. I then describe how the traditional three-state model is expanded to allow for alignment to

multiple variants at multiple independent positions in the reference sequence, which is the basis for the rest of this work.

### 2.1.1 Traditional Pair HMM

A pair-HMM with the structure shown in 2.1.1A was implemented.

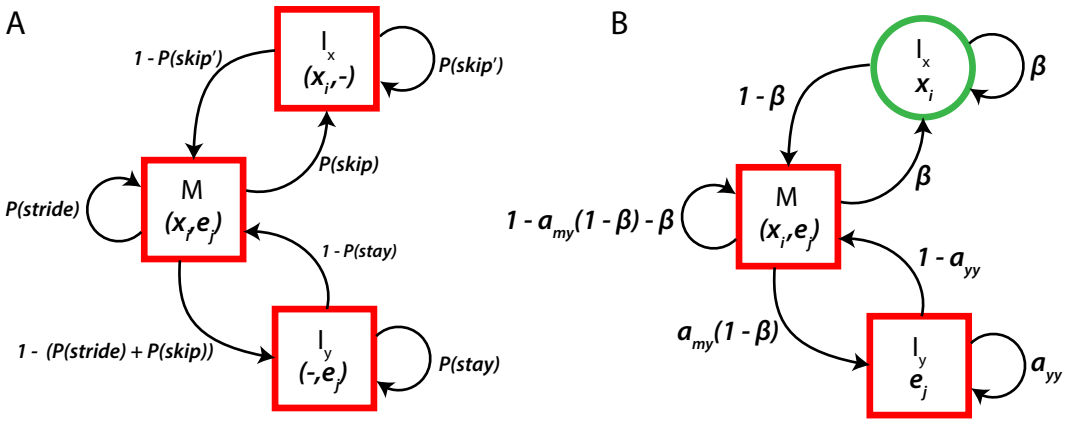


Figure 2.1: Pair-HMM and Conditional HMM based on Loman et al.

This model calculates the probability of an alignment $\pi$ given a sequence of events $E = \{e_0, \ldots, e_n\}$, a sequence of nucleotides, $S$, divided into nucleotide k-mers $S = \{k_0, \ldots, k_m\}$, and the model $\Theta$; $P(\pi|E, S, \Theta)$. The three states; match $M$ for matching one event with one nucleotide k-mer, insert-X, $I_x$, for pairing a nucleotide k-mer with a gap, and insert-Y, $I_y$, for pairing an event with a gap. The transition probabilities are initialized to naive estimates for the stride, skip, and stay probabilities that correspond to the enzyme advancing exactly one nucleotide, advancing more than one nucleotide,

and not advancing, respectively. In one version of the model, the emissions for the $M$ state and $I_y$ are the product of the probability of the ionic current mean and ionic current noise. We assume independence of the mean and noise variables, so the conditional probability of an event, $e_j$, for a given k-mer, $k_i$, is given by,

$$P(e_j|k_i) = P(\mu_j|k_i)P(\sigma'_j|k_i)$$

The mean and noise are modeled as a normal distributions $\boldsymbol{\mu_i} \sim \mathcal{N}(\mu_i, \sigma_i)$ and $\boldsymbol{\sigma_i} \sim \mathcal{N}(\mu'_i, \sigma'_i)$, where $\mu_i$, $\sigma_i$, $\mu'_i$, and $\sigma'_i$ are given by the lookup table (Section 3.1) or are the maximum likelihood estimate (MLE) (3.5). For alignments using the MLE we only update the $\mu_i$, $\sigma_i$ parameters. The $I_x$ is silent and does not emit. In another version of the model we modeled the ionic current distributions as a hierarchical Dirichlet process mixture of normal distributions (section 2.2.2). In this case we use the posterior mean density as the emission probability of the $M$, and $I_y$ states instead of the probability density function for a normal distribution.

### 2.1.2 Conditional Pair HMM

In the first iteration to improve the pair-HMM, a conditional model was used with similar underlying mechanics as the profile-HMM proposed by Loman et al. [33]. The basic architecture is shown in **??**B. This model calculates the probability of an event sequence given a nucleotide sequence, $P(e_1...e_n|k_1...k_n, \Theta)$ by using the given model for the individual k-mers, $\theta$, to calculate the probability of a k-mer being skipped due to it's similarity to the previous k-mer. Thus the probability of the transition from $M$ to

$I_x$ is now a function of the expected value of the mean for $k_i$ and $k_{i-1}$:

$$\delta = |\mathrm{E}(\mu_{k_i}) - \mathrm{E}(\mu_{k_{i-1}})|$$

$$\beta = f(\delta) = f(|\mathrm{E}(\mu_{k_i}) - \mathrm{E}(\mu_{k_{i-1}})|$$

The values of $\delta$ binned in 0.5 pA intervals. For illustrative purposes, suppose two identical adjacent k-mers such as a homopolymer run `TTTTTTT`, $\delta$ will be zero and the function will return the maximum allowed skip probability, $\underset{k}{\mathrm{argmax}} P(M_k \rightarrow I_x)$. On the other hand, two adjacent k-mers with drastically different expected mean current levels will return the lowest probability of transitioning to the skip state. The rest of the transitions are shown in 2.1.1. The two constant multipliers were adopted from Loman et al. and initialized without modification[1]. The emission probabilities for the $M$ and $I_y$ states are the same as the pair-HMM with the exception that the noise is modeled as an inverse Gaussian,

$$P_{\mu,\lambda}(x) = \left[\frac{\lambda}{2\pi x^3}\right] e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}$$

The $I_x$ state is silent therefore does not emit a event/k-mer pair.

### 2.1.3 Generalized HMM with Duration Modeling

Each event persists for a certain amount of time in between movements by the enzyme. The event duration, or *dwell time*, has historically been used to measure

[1] $p_{mx} = 0.17, p_{yy} = 0.55$

enzyme kinetics using nanopore sensors [31]. We sought to incorporate the duration of events into our model in order to account for potential errors in the segmenting algorithm and more accurately represent the biophysics of the nanopore. In the case where an event is split without the enzyme advancing the two resulting 'over segmented' events should be shorter than expected. Alternatively, if there is no noticeable change in current, for example in homopolymeric runs or kmers with similar expected current levels, the resulting 'under segmented' event should be longer than expected. The overall algorithm can be summarized,

$$P(e_j|x_i, x_{i+1}..., x_{i+n}) = \begin{cases} P(d_j|n)\frac{1}{n}\sum_{k=0}^{n-1} P(e_j|x_{k+i}), & \text{if } n > 0. \\ P(d_j|0), & \text{if } n = 0. \end{cases}$$

The structure of the HMM is shown in 2.1.3. The frequencies of the event durations were fit to a Poisson distribution

$$P(n|d_j) = Poisson(\lambda) = \frac{\lambda}{n!}e^{-\lambda}$$

where,

$$\lambda = cd_j.$$

The constant $c$ was estimated as shown in **??**. Then using Bayes theorem we can calculate the likelihood of $n$ for a given $\lambda$ by

$$P(\lambda|n) = B^{n+1}\frac{\lambda^n}{n!}e^{-2\lambda}$$

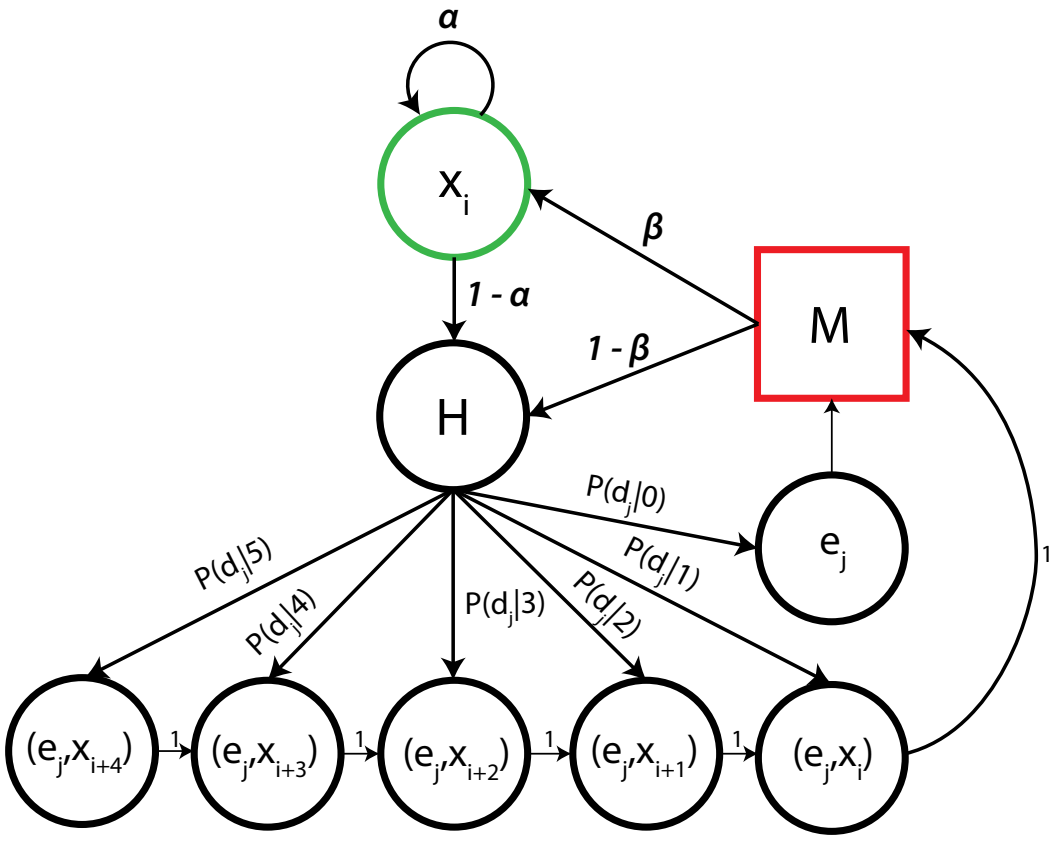


Figure 2.2: General HMM with Duration Modeling

Where $B$ is the rate parameter of the Gamma distribution defined as the conjugate prior to the Poisson. The calculation for the transition to the skip state, $\beta$, is identical as the previous model. The calculation for the skip self-loop, $\alpha$, follows

$$\alpha = f'(\delta) = f'(|\mathrm{E}(\mu_{k_i}) - \mathrm{E}(\mu_{k_{i-1}})|$$

with the exception that the bins for the $\alpha$ parameters are trained indepen-

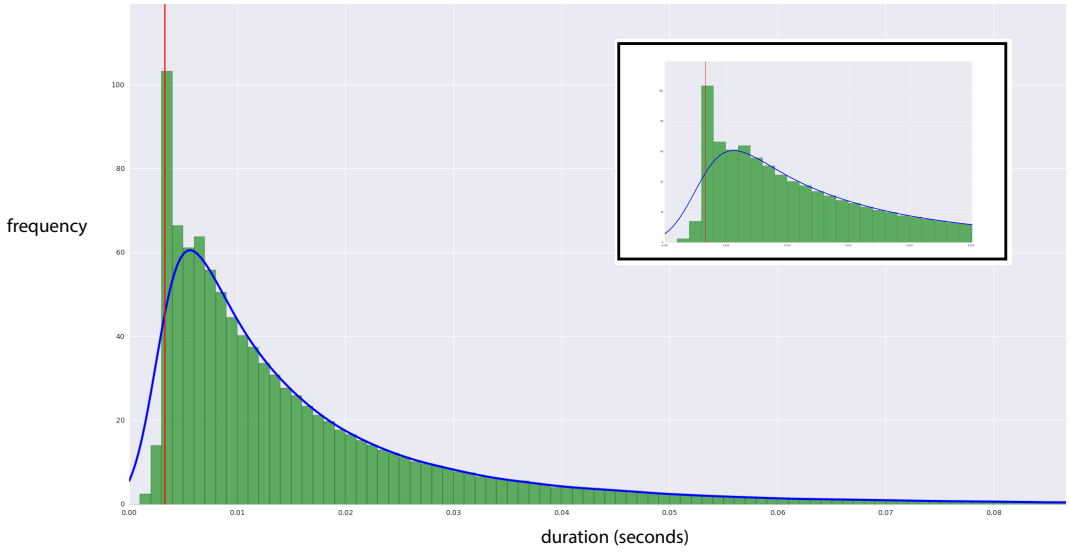


Figure 2.3: Plot of duration frequencies from sample of 17,000 events, $c$ estimated as the value at the red line.

dently from the ones for $\beta$. The emission probabilities for a given event/k-mer pair are calculated as before using the model $\theta$. For paths that involve matching one event to multiple k-mers the average of the probabilities for the k-mers is used. This was done as to not inappropriately 'penalize' matching events to multiple k-mers.

## 2.2 Hidden Markov model for variant calling

We model the signal with a variable-order pair HMM that tracks a reference sequence but allows a reference nucleotide to be any of several potentially modified bases Figure 2.2B. The model allows for multiple types of modifications to be mapped simultaneously at a single molecule level. We use hierarchical Dirichlet process mixture model to learn the effects different base modifications have on the ionic current 2.2 [49].

The HDP mixture model is a Bayesian nonparametric method that shares statistical strength to robustly estimate a set of related but potentially complex distributions.

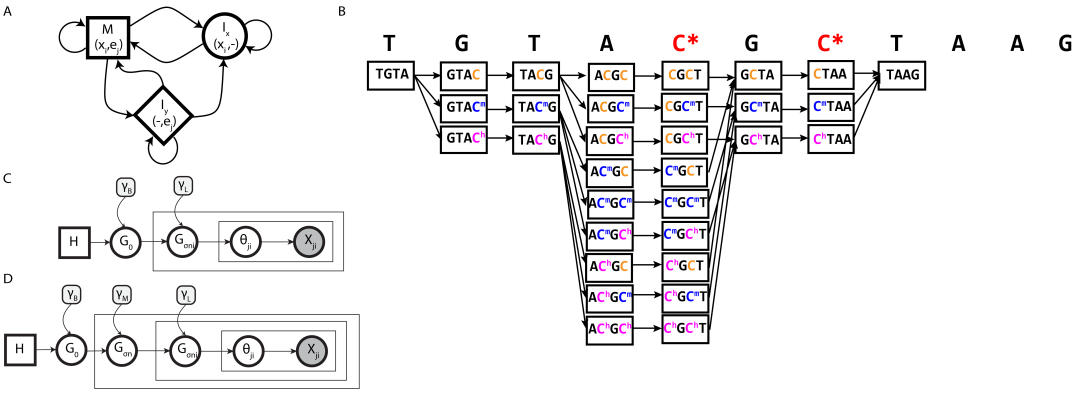


Figure 2.4: Overview of models. A. Architecture of hidden Markov model used in this work. The match state, M (square), emits an event-k-mer pair and proceeds along the reference and the event sequence, Insert-Y, $\mathbf{I_y}$ (diamond), emits a pair and proceeds along the event sequence but stays in place with respect to the reference, and Insert-X, $\mathbf{I_x}$ (circle), proceeds along the reference but does not emit a pair and stays in place with respect to the event sequence. B. Variable-order HMM meta-structure over an example reference sequence containing ambiguous methylation variants. Each $\mathbf{C}^*$ in the reference represents a potentially methylated cytosine. The structure expands around the $\mathbf{C}^*$ to accommodate all possible methylation states (in this case, C, 5-mC, and 5-hmC). Each cell contains the three states shown in A, and transitions span between cells. The transitions are restricted so that methylation states are labeled consistently within a path. The match states are drawn with 4-mers for simplicity, but the model is implemented with 5-mers and 6-mers. Two-level (C) and three-level (D) hierarchical Dirichlet process shown in graphical form. Circles represent random variables. The base distribution H is a normal inverse-gamma distribution for both models. The Dirichlet processes $\mathbf{G_0}$, $\mathbf{G_{\sigma n}}$, and $\mathbf{G_{\sigma ni}}$ are parameterized by their parent distribution and shared concentration parameters $\gamma\mathbf{B}$, $\gamma\mathbf{M}$, and $\gamma\mathbf{L}$. The factors $\theta_{\mathbf{ji}}$ specify the parameters of the normal distribution mixture component that generates observation $\mathbf{x_{ji}}$.

Our method calls methylation variants based on the posterior probability of aligned event to k-mer pairs. First, the model aligns the event sequence to all self-consistent k-mers containing each variant. The template and complement event sequences are aligned separately because the current distributions differ between the two

strands. We then marginalize over the HMMs states to obtain posterior probabilities for the methylation variants at a given base. We call the methylation status as the variant with the highest marginal posterior probability. To call methylation using multiple reads, we sum the probabilities from the individual reads aligned to a position and call the variant with the highest posterior mean.

### 2.2.1 Structure of variable-order hidden Markov model

Our HMM is structured to allow alignment of multiple different bases at a given position in the reference sequence. We term these positions *ambiguous positions*. Positions in the reference are flagged as ambiguous before the alignment begins. Ambiguous positions can be any subset of potential allowed nucleotides, for example $\mathbf{N} = \{A, C, G, T\}$ or $\mathbf{C} = \{C, mC, hmC\}$ for variant calling and 3-way cytosine classification, respectively. In three-way classification experiments on synthetic oligonucleotides (section 3.5), we allow for C, 5-mC, and 5-hmC to be aligned to a given cytosine. In two-way classification experiments (Sections 3.6, 3.7), the model is restricted to C and 5-mC or A and 6-mA in the cases of alignment to cytosine and adenine, respectively.

The fact that each event corresponds to multiple positions in the reference (Section 1.3, 2.2) means that more than one event reports on a single ambiguous position. Accordingly we model each event as reporting on a substring of nucleotides of length $k$, *k-mers*. When we allow for multiple variants in the reference sequence we would like to compute over all possibilities in a way that the probabilities for a given k-mer are tied

with only k-mers that share that particular variant pattern. As can be seen visually in Figure 2.2 and Figure 2.2.1, when a position is allowed to be variable (C* bases) the number of paths expands to accommodate the number of ambiguous positions.

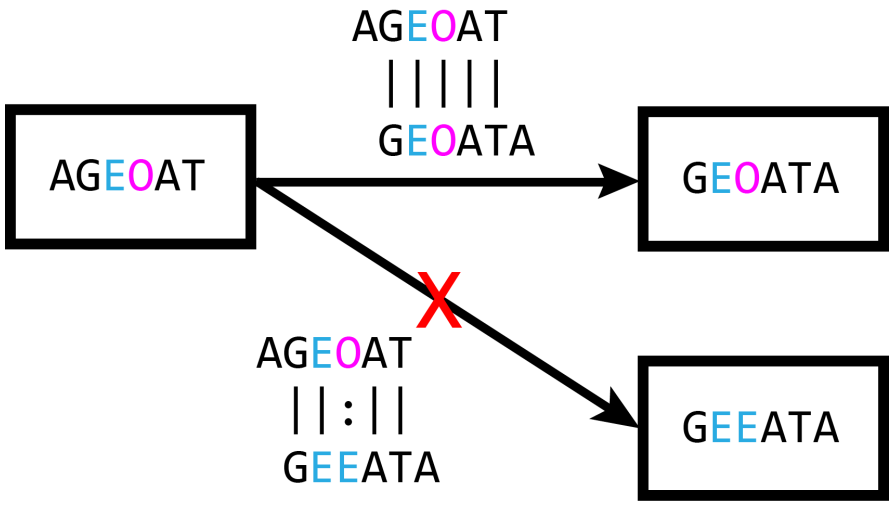


Figure 2.5: Transitions between cells in the dynamic programming matrix are only allowed between k-mers where the last k-1 bases of the first k-mer (AGEOAT) match the first k-1 bases of the next (GEOATA).

Given k-mer $k_i$ that contains $\eta$ variable bases within the set $\boldsymbol{B} = \{C, C^m, C^h\}$ (or $\boldsymbol{B} = \{A, A^m\}$, in the case of adenine methylation) the number of paths, $l$, is simply $l = \eta^{|\boldsymbol{B}|}$. The dynamic programming matrix is changed such that every cell has $l$ dimensions, which is precomputed based on the reference and the ambiguous positions. Then we perform the forward-backward algorithm through the matrix except that we don't want to sum over all paths, $\Pi$, only paths that represent legal moves, $\pi \subseteq \Pi$. A move is defined as legal if bases 2-k of the previous path's k-mer are the same as bases 1-(k-1) of the current path's k-mer. For example, assume 5m-C is represented at `E` and

5hm-C is O, the move between 6-mers AGEOAT and GEOATA would be legal, but the move between AGEOAT and GEEATA would not. Moves from the start state and to the end state are legal regardless of the k-mer. With this framework we can calculate the total probability of an event sequence, $E$, the reference sequence, $S$, containing ambiguous positions, and the model by:

$$P(E, S, \Theta) = \sum_{\pi} P(E, S, \pi)$$

Next we can calculate the posterior probability that 6-mer $x_i$ is aligned to event $e_j$ (noted as $x_i \diamond e_j$) as

$$P(x_i \diamond e_j | E, S, \Theta) = P(\pi_i = M | E, S, \Theta) = \frac{f_M(i) b_M(i)}{P(E, S, \Theta)}$$

Where $f_M$ and $b_M$ are the forward and backward variables respectively. The joint probability for the reads event sequence and the reference is calculated with the forward-backward algorithm, and the likelihood of each methylation variant at each ambiguous position is calculated by marginalizing over the HMMs states. To make methylation variant calls using multiple reads, the variant with the greatest posterior mean given the reads is used.

### 2.2.2 Hierarchical Dirichlet process mixture model

The hierarchical Dirichlet process (HDP) mixture is a statistical model in which a collection of mixture distributions are composed of a countably infinite set of shared mixture components. The weights of the components in each mixture distribution are

determined according to a separate Dirichlet process on the shared collection of components [49]. In addition, the mixture components themselves are distributed according to a Dirichlet process that draws components from a base distribution.

We model the distribution of ionic currents across the different k-mers as a hierarchical Dirichlet process mixture of normal distributions. In this model, each current distribution is composed of a countably infinite collection of Gaussian mixture components that are shared between the k-mers. More precisely, all of the distributions draw mixture components according to a Dirichlet process over the same discrete distribution over a countably infinite collection of "atoms", each of which consists of the parameters for a normal distribution. In addition, this discrete distribution is itself generated according a Dirichlet process on the normal-inverse-gamma distribution (a conjugate prior to a normal distribution, that is to the mixture components). Sharing mixture components statistically shrinks our estimates of the current distributions toward each other. This boosts statistical strength since each distribution can share the information learned by the others. We also have the option of adding a further layer of Dirichlet processes between the Dirichlet process that generates the distribution over shared components and the Dirichlet processes that generate the k-mer distributions. After doing so, the Dirichlet processes are arranged in a tree structure Figure 2.2C and D. This encourages a greater degree of shrinkage within each subtree. We experimented with several topologies for this tree, each representing a different grouping of k-mers based on their sequence composition (see 2.2.3 for results).

The most intuitive interpretation of the Dirichlet process for this setting is the

"stick-breaking procedure". In this construction, we draw a countably infinite number of atoms from the normal-inverse-gamma distribution and then form a new distribution over these atoms by assigning them weights $w_k, k = 1, \ldots, \infty$ sequentially:

$$w_k = \beta_k \prod_{i=1}^{k-1}(1 - \beta_i) \qquad \beta_i \sim \text{Beta}(1, \alpha) \tag{2.1}$$

where $\alpha$ is a hyper-parameter referred to as the "concentration". Note that both the value of the atoms and their weights are random variables. To generate each of the current distributions, the HDP draws a countably infinite set of mixture components according to a Dirichlet process on this new distribution (treating the weights as probability masses). The weights that this process assigns to the atoms are the weights on the components in the mixture distribution.

Our motivation for using an HDP to model signal distributions is that it shrinks the set of distributions it learns towards each other, which increases robustness, while retaining the flexibility to approximate any arbitrary distribution given sufficient data (allows for each k-mer to have a potentially complex emission distribution). Both of these features are important, since we expect that the current distributions may have complex shapes and we must estimate a large number of them. The shrinkage is a result of the fact that all of the distributions share the same mixture components: each distribution can share the information learned by the others. In addition, this allows us to calculate informed prior distributions even for k-mers that have not been positively identified in the training data, a feature that will be useful for expanding the scope of

modifications that the model can detect.

### 2.2.3 Grouping 6-mers with different HDP topologies

Since the biophysical relationship between each given k-mer sequence and the observed ionic current distribution is poorly understood, we empirically tested whether different subgroupings would increase statistical strength using reads from the synthetic oligonucleotides. We tested HDP models with five different subgroupings of 6-mers, and compared their single-molecule accuracy on the synthetic oligos. The two-level HDP does not separate them into any subgroups Figure 2.2C, whereas the rest of the models group 6-mers by features of their 6-mer sequence Figure 2.2D.

In the original HDP described by Teh, et al. (2006) [50], all of the mixture distributions draw mixture components from the same distribution over atoms, which is generated by a Dirichlet process on the base distribution. However, there is a relatively simple extension of this model in which the mixture distributions are split into predetermined groups, and it is these groups that share a distribution of over atoms. This is accomplished by adding an additional layer of Dirichlet processes between the one that generates the initial distribution over atoms and the ones that generate the mixture distributions. To conceptualize this, it helps to think of the HDP as a collection of Dirichlet processes arranged in a tree structure (with the Dirichlet process over the base distribution at the root). With this framing, the generalization amounts to having a tree with a depth of two instead of one.

All of the mixture distributions still share the same collection of atoms, since

all of the atoms in the middle layer of Dirichlet processes are drawn from the original root Dirichlet process. However, the weights of the atoms are reassigned according to a new stick-breaking procedure (See Equation (2.1)) in each of the middle-level Dirichlet processes. The effect is that the shrinkage between the distribution estimates is greater within a subtree than between subtrees. Since we have control over the topology of the tree, this serves an extra "knob" that can be used to increase statistical strength. However, the grouping of mixture distributions into subtrees presumably must reflect clusters of similarity in the true distributions in order to accomplish this goal.

We took an empirical approach to determining what topologies for the tree of Dirichlet processes would be meaningful. We came up with several ways to partition 6-mers based on their sequence composition and tested the performance of each one. The groupings were as follows:

1. *No groups*: this model has no middle layer of Dirichlet processes.

2. *Groups of 6-mers containing the same number of purines and pyrimidines*: this corresponds to a hypothesis that steric bulk is a strong determinant of the current distribution.

3. *Groups of 6-mers that shared the same two middle nucleotides*: this corresponds to a hypothesis that the nucleotides that are passing through the most constricted portion of the nanopore have the most influence.

4. *Groups of 6-mers that contain the same nucleotides, irrespective of their order*: this corresponds to a hypothesis that the position of nucleotide in the 6-mer is

less important than which nucleotide it is, similar to resistors in series.

5. *Groups of 6-mers that contain the same nucleotides, irrespective of both their order and their modification status*: this corresponds to the same hypothesis as the previous grouping and also a hypothesis that the modification status has relatively little effect on the current distribution.

## 2.3 Supervised training of model parameters

We train the HMM with a variant of the Baum-Welch procedure. First, we heuristically initialize the emission distributions by training them on aligned events above a probability threshold (0.9 for the synthetic oligonucleotides and 0.8 for the E. coli and plasmid DNA) from the preliminary alignment described in Section 3.2. In the three-way classification control experiments on synthetic oligonucleotides using normal distributions, this entails calculating the maximum likelihood normal distribution for each 6-mer. For the HMM-HDP, we estimate the posterior mean density for each k-mers distribution using a Markov chain Monte Carlo (MCMC) algorithm (Section 2.4). In both cases, only the distributions for the ionic current means are learned following the preliminary alignment. A separate neural net experiment suggested that the event noise did not add to classification accuracy (Section 3.4). At this step, we also re-estimate the HMMs transition probabilities independently. For experiments on synthetic oligonucleotides, training was performed in batches of 15,000 nucleotides and iterated 20 times. For experiments on *E. coli* and pUC18 reads, batches of 30,000 bases and 30 iterations

were performed. When experimenting with different emissions models (eg. different HDP topologies, Section 2.2.3), the transition matrices were trained specifically for that model. We then produce new alignments and re-estimate the emission distributions from high confidence assignments as in the initialization. This process is iterated until the models variant calling accuracy stops improving. The MCMC algorithm we use for the HDP is the Chinese Restaurant Franchise Algorithm, a Gibbs sampler for HDP mixture models (Section 2.2.2). We discard the first 30-times the total number of assignment data points as burn-in and collect 10,000 samples, thinning sampling iterations by 100. Whenever we record samples from the Markov chain, we evaluate the posterior predictive distribution for each 6-mer at a grid of 1200 evenly spaced points in the interval between 30 pA and 90 pA for R7.3. For R9 experiments we use a grid of 1800 evenly spaced points in the interval between 50 pA and 140 pA. After sampling, we compute our estimate of the posterior mean density as the mean of the sampled densities at each grid point. Subsequently, we interpolate within the grid using natural cubic splines.

## 2.4 Training the HMM-HDP model

We train the HMM-HDP by an iterative expectation maximization and Gibbs sampling procedure. However, this method is sensitive to the initial values of the model's parameters, so first we leverage a standard statistical model to heuristically initialize the HMM's emission distributions to normal distributions. We use a lookup table provided by ONT that consists of parametric distributions that describe the events arising

from each of k-mers composed of the four conventional nucleotides, (Section 2.6.3). To generate preliminary alignments we used the parameters lookup table to calculate the probability an event being due to a particular k-mer in the Match and Insert-Y states of the HMM. The events' mean current and fluctuation in the mean (noise) are modeled as normal distributions. We assume independence of the mean and noise variables, so the conditional probability of an event for a given k-mer is just the product of the mean and noise marginal probabilities (section 2.1).

This initial HMM only has emissions for the canonical four-nucleotide alphabet. This allows us to use the standard statistical model and use a first-order HMM, described in Section 2.1.1 since there are no ambiguities between the HMM's alphabet and the reference alphabet. We use the banded alignment scheme described in Section 3.2 to obtain a posterior probabilities that an event was generated by a given k-mer. Our experimental design allows us to then label the methylation status of the cytosines (or adenines) from these alignments *post hoc*. After doing so, we extract the aligned events with posterior probabilities of at least 0.9 (0.8 for *E. coli* and pUC19 alignments) and use these as training data to learn emission distributions for the HDP.

Once we have a set of aligned events as training data, we estimate the emission distributions using an MCMC method described in section 2.3. Briefly, it involves integrating out the values of the HDP's atoms and then sampling the full conditional distributions of latent variables that indicates which mixture component generated each data point [50]. For models trained on *E. coli* we record 15,000 samples. Every time we record a sample, we compute the posterior predictive distribution of each k-mer on

a grid that covers the range of MinION current signal. We then estimate the mean posterior density at each of these grid points from the sample and interpolate between them using natural cubic splines. These serve as our distribution estimates.

We then proceed in a similar fashion to the heuristic initialization. First, we align the reads to the reference, except now we use the HMM with the emission distributions estimated by the HDP. We then extract the assignments with posterior probabilities greater than or equal to the threshold for that dataset and use these as training data to obtain mean posterior distributions from the HDP. We then iterate this process until classification accuracy stops improving. This differs from the true Baum-Welch procedure since we are using posterior mean estimates rather than *maximum a posteriori*. However, these values are asymptotically equivalent in unimodal posteriors, so this is probably a reasonable approximation.

## 2.5 Computing posterior probabilities of alignments and ungapped alignment scores

To have an empirical measure of the quality of an alignment (and its underlying event sequence) we use an average of the posterior match probabilities from the alignment. As mentioned previously, the HMM takes the event sequences as input and aligns them to a reference nucleotide sequence. Let $x_i$ be a k-mer in the reference sequence $S$, and $e_j$ be an event in the event sequence $E$, and $e_j \diamond x_i$ mean that event $ej$ is aligned to k-mer $x_i$. The model calculates $P(x_i \diamond e_j | E, S, \Theta)$, the posterior probability

for event/k-mer aligned pairs given the model $\Theta$. From here we calculate the ungapped alignment score, U.A.S., which is defined as

$$U.A.S = \frac{P(x_i \diamond e_j | E, S, \Theta)}{N} \tag{2.2}$$

where N is the total number of aligned pairs in the alignment.

## 2.6 Variant calling and error correction

This section describes a method that probabilistically identifies variation between a known nucleotide reference sequence (for example, an assembly or candidate reference sequence) and a latent reference sequence (the 'correct' or true sequence, $S^*$) that generated the observed reads. The HMM is used as the generative model, $\theta$, for all sequences, $\mathbf{S}$, that contains the desired sequence which generates the observed sequencing reads, $\mathbf{R}$, (Figure **??**). Where $\mathbf{S}$ is the set of all possible sequences $\{S_i, S_{i+1}, ..., S_k\}$ of length $l$. Finding the sequence $S^*$ that satisfies

$$S^* = \underset{S \in \mathbf{S}}{\operatorname{argmax}}\ P(S|\mathbf{R}) = P(\mathbf{R}|S)P(S|\theta)$$

could potentially involve exploring all $4^l$ possible sequences. However, instead of starting with a random reference sequence the reference sequence is initialized by mapping the sequencing reads to a known reference ($S^I$) or assembly. Then it can be assumed that the majority of the bases in $S^I$ will be in $S^*$. The two extreme examples would be in the case of sequence error-correction where 10-25% of bases will be errors

(mismatches between $S^I$ and $S^*$), on the other hand, in variant calling, <1% of the bases might be mismatches.

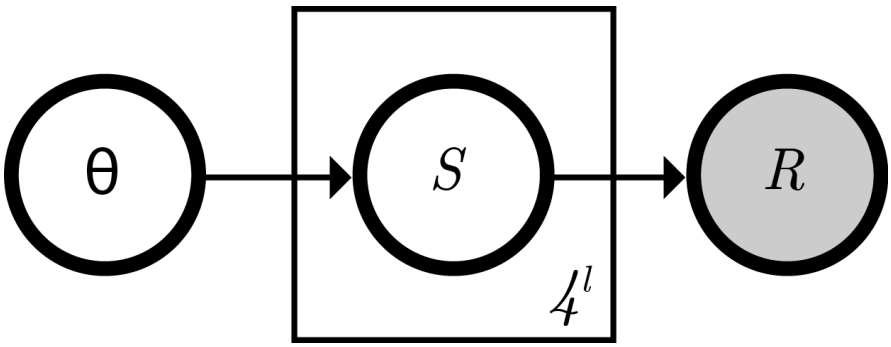


Figure 2.6: Generative model for all sequences of length $l$ composed of the canonical 4 nucleotides.

The total number of potential sequences might still be very large. Distant bases in the reference can be assumed to be independent given the data due to a reasonable assumption of linkage disequilibrium, so multiple bases can be tested at once effectively parallelizing the search process. The HMM model also simultaneously aligns all bases $\{A, C, G, T\}$ to an ambiguous position, `N`. To further constrain the search space, instead of checking for edits at every base, only a subset of the bases are checked.

First a method is described that proposes bases to be checked for edits (changing base identity). Followed by two methods for exploring the posterior probability distribution over the possible sequences within the edit distance of $S^I$ given the proposed edits.

### 2.6.1 Proposing edit positions

#### Notation

The initial (input) reference sequence $S^*$ of length $l$ contains bases from the set $\mathcal{N} = \{A, C, G, T\}$ to afford a sequence $\{s_0, s_1, \ldots, s_l\}$ of nucleotides. The marginal probability of a base in $S$ having a particular identity, denoted $s_i \leftarrow n$ where $n \in \mathcal{N}$, for a given read, $r$, is calculated by summing the posterior probability of all events aligned to a k-mer where the appropriate base in the k-mer is $n$.

$$P(x_i \leftarrow n|r, \theta) = \frac{\sum_{i|k_{s_i}=n} P(e_j \diamond k_i|\theta, r)}{\sum_i P(e_j \diamond k_i|\theta, r)}$$

The posteriors for all of the reads are summed and normalized, to give the final marginal posterior probability.

$$P(x_i \leftarrow n|\mathbf{R}, \theta) = \frac{\sum_r^{|R|} P(x_i \leftarrow n|\theta, r)}{\sum_r^{|R|} P(x_i|\theta, r)}$$

### 2.6.2 Scanning all bases to propose edits

The first problem in error-correction/variant calling is to determine the bases that should be checked for edits without having to check all possible sequences. A function called `ScanForProposals` (Algorithm 1) takes as input a reference nucleotide

sequence $S = \{s_0, s_1, \ldots, s_l\}$ and a set of ionic current event reads, $\mathbf{R}$. It outputs **Proposals**, $C$, a list tuples containing base position, and the difference between the marginal posterior probability for the base in the current reference, $n$, and an alternative base, $n'$, $(\nabla_i = P(s_i \leftarrow n'|\mathbf{R}, \theta) - P(s_i \leftarrow n|\mathbf{R}, \theta))$.

$$C = \{(x_{\{l\}}, \nabla_{\{l\}}), \ldots, (x_{\{l\}}, \nabla_{\{l\}})\}$$

$$\text{where } \{l\} \text{ denotes } l \in \{0, \ldots, len(S)\}$$

To begin a set of sequences $\mathcal{S}$ is generated where every *step* bases are changed to `N` for every register $<$*step*. Below is an example where *step* $= 6$.

$S$ = AGAATTGGTTAATTGGTT

$\mathcal{S}_0$ = NGAATTNGTTAANTGGTT

$\mathcal{S}_1$ = ANAATTGNTTAATNGGTT

$\mathcal{S}_2$ = AGNATTGGNTAATTNGTT

...

$\mathcal{S}_5$ = AGAATNGGTTANTTGGTN

Every read is aligned to every sequence in $\mathcal{S}$ in parallel (Algorithm 1).

### 2.6.3 Estimating the posterior distribution of single-base edits

After determining positions where there is evidence that a base in the reference should be edited, (2.6.2), the next step is to estimate the posterior distribution of the bases at these positions. The MinION reads words of DNA, so neighboring positions within a certain window cannot be considered independent. Accordingly, the function

INITIALIZE $C \leftarrow \{\}$

**for** $s_i$ in $S$ **do**

  **if** $s_i' = \underset{n}{\text{argmax}}\ P(s_i \leftarrow n|\mathbf{R}, \theta) \neq s_i$ **then**

    Add $(x_i, \nabla_i)$ to $C$

  **end if**

**end for**

**return** $C$

**Algorithm 1:** `ScanForProposals`

`GroupSitesInWindow` groups the elements in $C$ if they are within a window. Elements in $C$ that aren't grouped together can be evaluated independently so the inner loop can actually be evaluated in parallel. The length of *window* can be anything, but empirically $\geq 6$ works well. This turns $C$ into a set of lists of tuples of co-localized elements.

### 2.6.4 A method based on gradient descent

This method attempts to iteratively refine the reference sequence by prioritizing the proposal positions with higher $\nabla$. As input the function takes the initial reference sequence, $S^I$, and the set of lists of proposals, $C$. It outputs a set of key-value pairs, where the key is the position in the reference sequence and the value is an array containing the frequency of each base at that position, denoted $F$.

From each list in $C$ take the element with the maximum $\nabla_i$ and change those positions in $S$ to `N` (the wild card character that will allow the model to calculate the probability of each base in $\mathcal{N}$ of being at that position). The positions are then

changed to the base with maximum posterior probability, yielding $S'$, and the sequence is re-scanned for proposal positions. The iteration continues until there are no more proposals or a set number of *cycles* is reached. The full algorithm is described in detail in Algorithm 2.

1: INITIALIZE $S \leftarrow S^I$, $C \leftarrow$ `GroupSitesInWindow`(`ScanForProposals`($S^I$))
2: **for** *cycle* in $0 \ldots cycles$ **do**
3:  **for** $s_i$ in $\underset{\nabla}{\text{argmax}}\ C$ **do**
4:  Align all reads in $\mathbf{R}$
5:  $s_i' \leftarrow \underset{n}{\text{argmax}}\ P(s_i \leftarrow n | \mathbf{R}, \theta)$
6:  $F[i][n]$ += 1
7:  **end for**
8:  $S \leftarrow S'$
9:  $C =$ `GroupSitesInWindow`(`ScanForProposals`($S'$))
10:  **if** $C = \{\}$ **then**
11:  break
12:  **end if**
13: **end for**
14: **return** $F$

**Algorithm 2:** `GD`

### 2.6.5 A method based on Gibbs Sampling

This method does not prioritize proposal positions based on $\nabla$, instead it randomly picks from each group in $C$ and records the frequency of each nucleotide in $F$. It performs the iteration for a given number of *sweeps*. The input and output are the same as the above function. Described in Algorithm 3.

1: INITIALIZE $S \leftarrow S^I$, $C \leftarrow$ `GroupSitesInWindow(ScanForProposals(`$S^I$`))`
2: **for** *sweep* in $0 \ldots sweeps$ **do**
3: **s** = Sample $s_i$ **i.i.d** from *groups* in $C$
4: **for** $s_i$ in **s do**
5: Align all reads in **R**
6: $F[i][n]$ += $\underset{n}{\text{argmax}}\ P(s_i \leftarrow n|\mathbf{R}, \theta)$
7: **end for**
8: **end for**
9: **return** $F$

**Algorithm 3:** `GibbsSampler`

### 2.6.6 Generalizing the HMM-HDP model

Nanopore sequencing has the unique feature that the sensor actually touches the DNA as it's being sequenced. As we've shown in this study, the waveform of ionic current can be mined for more information than just the canonical base identities. While ground-truth training data can be generated for some biologically-relevant DNA modifications (5-mC at CpGs, for example) by using enzymatic treatment or synthetic oligos,

*de novo* discovery of new modifications is a more difficult problem. The entry point to any study going after new or not completely characterized modifications (using the MinION) will be analysis of the ionic current. The HDP has a convenient ability of adding an amount of regularization when learning new k-mer distributions that makes it particularly adept to this task. One could imagine testing a hypothesis where methylation is suspected at a motif due to the presence of a methyltransferase gene. Changes due to methylation can easily be investigated by using our naive HMM (Figure 3), which can be followed up by approximate labeling of the motifs in question and training the HDP to detect these modifications. This form of bootstrapping will likely not be as accurate as a model trained with perfectly labeled training data. However, the relative ease and low equipment cost make the combination of MinION sequencing with our method an appealing research avenue for investigating DNA modification.

## 2.7 Scaling methylation calling and alignment pipelines with cloud-based workflows

To handle the large data volume, the original `marginAlign` [22] and `SignalAlign` [42] algorithms were adapted to cloud infrastructures using the Toil batch system (Vivian *et al. BioRxiv* 2016). Toil allows for computational resources to be scaled horizontally and vertically as a given experiment requires and enables researchers to perform their own experiments in identical conditions. Workflow diagrams are shown in Figure 2.7.

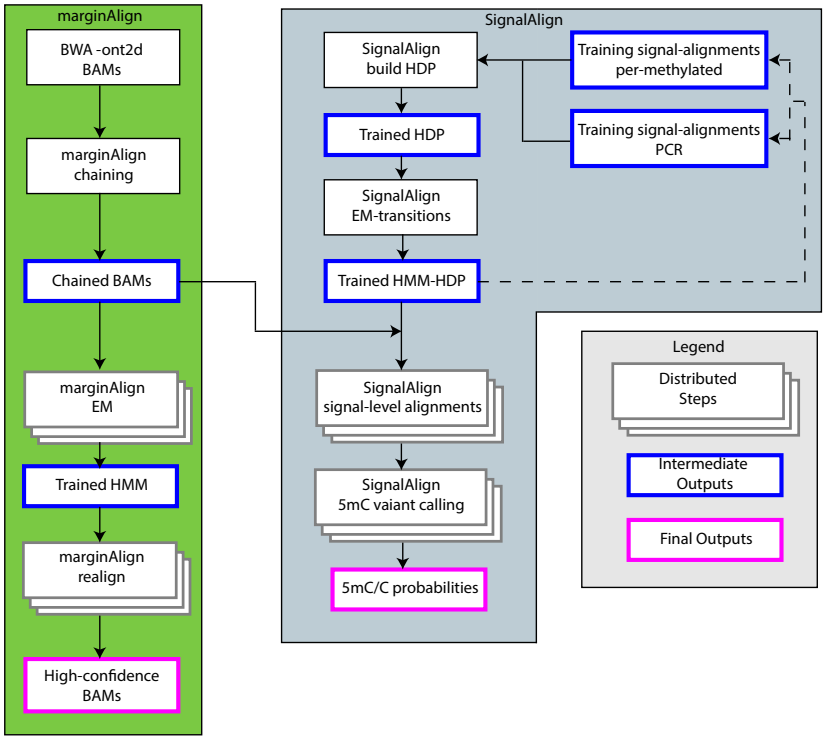


Figure 2.7: Cloud-based workflow showing how analysis can be scaled horizontally to consume more input data.

# Chapter 3

# Detection of chemical modifications in genomic DNA

### Publication Note

The results in this chapter were published here: [42]. The results were obtained in collaboration with my co-authors and the words herein are therefore influenced by them, however they have been adapted for this thesis. Section 3.8 contains unpublished results that were obtained in collaboration with the co-authors of *Jain et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. BiorXiv April 2017* a manuscript is currently under review.

## 3.1 Estimating emission distributions for R9 nanopores

For the R7.3 sequencing protocol, and earlier versions, ONT provided a lookup table describing the expected normal distribution of ionic currents for each k-mer. In

all experiments with the R7.3 chemistry, we used this table to initialize the HMM's emission distributions. However, with the release of the R9 chemistry, ONT switched to a new base calling algorithm that no longer utilizes the lookup table and they no longer provide it with sequencing reads. We obtained an R9 lookup table directly from ONT, but it described distributions for 6-mers rather than the 5-mers that R9 model uses.

We estimated a new set of distributions to serve as a 5-mer lookup table for the R9 chemistry. Doing so is complicated by the fact that we need to use normalized events for the estimates to make them generalizable. However, before estimating the distributions, we do not yet have distributional expectations with which to normalize the reads. To break this circularity, we first computed a set of heuristic normalization parameters based on the R9 6-mer table (privided by ONT for the R9 chemistry), and then used these to estimate the distributions in the 5-mer table.

Our estimates were based on a set of 1D reads from a methyltransferase deficient (*dam-/dcm-*) *E. coli* (NEB cat. no. C2925I) whole genome sequencing run. First, we obtained a set of 6-mer calls from the table we obtained from ONT anchoring the calls at the positions of the 5-mer calls in the reads. We then used the 6-mer calls to estimate the normalization parameters for our reads. We estimated the scale, shift, and drift parameters for each of our reads using a weighted linear regression.

$$e_i = SC \cdot \mu_i + DR \cdot t_i + SH + \epsilon \tag{3.1}$$

where $e_i$ is the current of the $i$-th event, $t_i$ is its time from the start of the experiment,

and $\mu_i$ is the mean for the corresponding 6-mer in the lookup table. The weights for each event were $1/\sigma_i^2$, where $\sigma_i^2$ is the 6-mer's variance. Technically, the weights should have also included the variance normalization parameter, but this will not result in any bias, only some inefficiency. We then estimated the variance normalization parameter for each read by measuring the overdispersion after accounting for the other normalization parameters.

$$VAR = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(e_i - SC \cdot \mu_i - DR \cdot t_i - SH)^2}{\sigma_i^2}} \tag{3.2}$$

Next we used the 5-mer calls from ONT's software and the normalization parameters to compute a new 5-mer table. First, we normalized each event.

$$\tilde{e}_i = \frac{e_i - DR \cdot t_i - SH}{SC}. \tag{3.3}$$

We then use a weighted linear regression to estimate the means for each 5-mer.

$$\tilde{e}_i = \mu_i + \epsilon \tag{3.4}$$

where we are now treating $\mu_i$ as a coefficient on an indicator variable for the 5-mer call. The weights in this regression are $1/(VAR^2 \cdot \tau_i^2)$ where $\tau_i$ is the standard deviation of the current signal for event $i$. Finally, we estimate the variance of each 5-mer's normal distribution.

$$\sigma_i^2 = \frac{1}{N-1} \sum_{j=1}^{N_i} \frac{(e_{i_j} - \mu_i)^2}{VAR_{i_j}^2} \tag{3.5}$$

## 3.2 Mapping of reads and event alignment

Alignment of the ionic current events from each read to the reference sequence is a two step process. Briefly, first we generate a guide alignment between the reads' nucleotide sequence and the reference, which we then use to constrain a second alignment of events to the reference. The guide alignment uses a concatenated sequence from Metrichors 2D alignment (Section 3.3) table, which allows for each base to be mapped to an event in the template and complement event sequence. This nucleotide sequence is aligned to the reference with BWA-MEM in `ont2d` mode [30]. We refer to this as the *guide alignment*. We use a banded alignment heuristic to increase the speed and decrease memory requirement of our HMM algorithm. The banding procedure is described in detail in [40]. Briefly, from the guide alignment runs of un-gapped matches are used as constraints in the edit graph around which we compute dynamic programming bands. To prevent any edge effects, the constraints are trimmed at either end by 14 nucleotide pairs. We then expand around the constraints to form the band by 50 anti-diagonal cells. To increase the efficiency of the algorithm, we break the alignment into fragments. Alignment bands computed between the centers of the constraints where the quality of the guide alignment should be highest. This allows our higher order HMM to have a smaller memory footprint as well as constraining the posterior probability distribution

to higher likelihood matches.

## 3.3 Making a read sequence from the 2D alignment table

The current MinION DNA sequencing library preparation protocol involves ligating a DNA hairpin to the distal end of the DNA duplex to be sequenced. This effectively makes the sequencing substrate one long strand of nucleic acid polymer. The sequencer then proceeds to sequence both strands of the DNA duplex. When these two "1D" reads are base called there is an event attributed to each nucleotide in the read. In the case of stays (where multiple events can be mapped to a k-mer) we take the pair with the highest probability (`p_model`). For skips (where there is no event for a base), we assign the bases to the previously matched event. The two 1D reads are assembled *in silio* into a "2D" read. During the assembly process some bases are inferred by the algorithm and do not have events attributed to them, as they are likely due to a combination of the two 1D sequence k-mers. We need every nucleotide in the sequence to correspond to an event because we use BWA-MEM to map the nucleotide reads to the reference sequence and use runs of consecutive matches to constrain the dynamic programming (see 3.2). Obviously, the inferred bases cause problems when we try to map bases to events. The "2D Alignment" table, however, does attribute each nucleotide position to an event. Therefore, we use the list of k-mers in the 2D alignment table to construct our nucleotide sequence, which is then used for when mapping the reads to the reference and generating the anchors.

## 3.4 Classification of ionic current events with neural networks

We investigated the feasibility of cytosine methylation detection by testing whether or not events aligned to a single cytosine could be classified based on their methylation status. We also used this strategy to evaluate which features are the most discriminatory in classification. Artificial neural networks are non-sparse classifiers well suited to this task, in this section we describe classification of events that have been aligned to the reference sequence using the preliminary alignments generated without consideration for methylation status.

### 3.4.1 Data processing for single cytosine motifs in synthetic oligonucleotides

Subsequences of the reference sequence, $motifs$, were selected that contain a single cytosine among a run of 11 nucleotides. The MinION reads both strands of the DNA duplex and combines these reads into a 2D, high quality read. The motifs we chose contain guanine bases, however, and the complement would therefore contain a variable number of modified cytosine bases and contribute to the classification accuracy. We therefore classified only one strand per read. In the case of forward-mapped reads this was the template strand, in the case of backward-mapped it was the complement strand. For this analysis, we only used data from the barcoded reads for training and testing.

To assess the amount of classification bias due to the stands alone, we classified *null* sites that do not contain any cytosines. The features of the *null* motifs should be the same between strands and the classifier should report accuracy close to random chance (33%).

We used the methylation-naive HMM (Section 2.3) to align events to the reference sequence. Events that aligned to positions in the reference corresponding to a motif were culled. When multiple events were aligned to a position, the one with the highest posterior probability was taken. Thus an ordered set of a maximum of 6 observed event mean current levels, six observed ionic current noise levels, and six posterior probabilities, was obtained. The difference from the observed mean current level from the expected current level (from the ONT table) given the 6-mer was taken for each observation. The same was done for the noise levels. The differences in the mean current level ($\Delta\mu$), differences in noise level ($\Delta\sigma$), and the posterior probabilities ($P$) became the features input to the classifier. We experimented with four different feature sets; $\Delta\mu$ alone, $\Delta\mu$ and $\Delta\sigma$, $\Delta\mu$ and $P$, and $\Delta\mu$, $\Delta\sigma$ and $P$.

### 3.4.2 Network architecture and training routine

Classification of the feature vectors was performed using a custom artificial neural network implemented in Theano [4]. For the individual motif classification a network with two hidden layers and a final softmax layer was used. The dimensions of this network were 50, 10, 3, with rectified linear unit and hyperbolic tangent non-linearity activation functions used for the first and second hidden layers respectively.

For classification, the class with the highest probability from the softmax layer was chosen. We split the dataset into three groups; 10% of the data was held out for testing after the training procedure. Of the remaining, training data, 50% was used as cross-training (validation) and 50% was used in the optimization. An equal number of feature vectors for each category were used in all data sets. Training of the network was done using mini-batch stochastic gradient descent with an annealing learning rate schedule. We used 5 reads per batch and a dynamic learning rate initialized at 0.1%. With decreases in cross-training batch costs, the learning rate was decreased by 10%, if the cross-training batch cost increased, the learning rate was increased by 5%. We used L1 and L2 regularization of 0.01. Lastly, the data was centered and normalized based on the training data set before starting the routine. The model that had the highest cross-training accuracy during the learning process was used for final evaluation of the test set. We performed the same training routine on the *null* motifs.

### 3.4.3 Classification accuracy is maximized using $\Delta\mu$ and posterior probability as features

The accuracy for the different feature sets for the cytosine-containing $motifs$ is summarized in Table 3.4.3, the *null* motifs are summarized in 3.4.3. The highest accuracy was obtained when using the mean and posterior probability at 65% on the template reads and 66% on complement reads. Including $\Delta\sigma$ did not appear to increase the accuracy of the classifier. We classified events culled from alignments generated with the HMM-HDP and the methylation-naive HMM. Both data sets produced similar

error rates when classified with the neural network.

## 3.5 Single molecule discrimination between C, 5mC, and 5-hmC on synthetic oligonucleotides

We performed three-way classification experiments between three cytosine methylation variants with synthetic oligonucleotides, each bearing a type of cytosine methylation variant. The DNA substrates are 897 base pairs long, and they contain 201 and 246 cytosines on the forward and reverse strands respectively. These experiments were performed with the R7.3 sequencing chemistry, so the k-mers were modeled as 6-mers. The reads span the full length of the substrate. The cytosines are distributed such that the model needed to learn 2868 new 6-mer distributions with methylated bases in addition to the 1784 canonical 6-mers contained in the forward and reverse reference sequences. We treat the two strands as independent, so a given event sequence is aligned to the appropriate strand and reports on only the cytosines in that strand.

We measured the per-read accuracy by the proportion of correct methylation calls on a single strand from a read. The best performing model was the "Multiset" model 3.5. However, it was a small gain in accuracy over the simpler ungrouped model. Based on these results we used the "Multiset" model for all further analyses. The mean and median per-read accuracy were 74% and 80% respectively for the template reads and 67% and 76% for the complement reads. The distribution of per-read accuracies is shown in Figure 3.5A. These results represent a significant improvement over the 33%

Table 3.1: Classification of *null* motifs

| | Forward/Template | | | |
|---|---|---|---|---|
| Motif Sequence | $\{\Delta\mu\}$ | $\{\Delta\mu, \Delta\sigma\}$ | $\{\Delta\mu, P\}$ | $\{\Delta\mu, \Delta\sigma, P\}$ |
| TTGTTGAATAA | 40 | 40 | 43.33 | 43.33 |
| GAGTTGAAGGA | 38.33 | 41.67 | 46.67 | 41.67 |
| GGATGATGGGG | 40 | 40 | 35 | 35 |
| AGGGGTAAAAG | 40 | 40 | 60 | 38.33 |
| AGGATGAAGGT | 40 | 40 | 40 | 40 |
| GAGGAAGGTGA | 40 | 40 | 40 | 40 |
| AAAAGAGTTTG | 40 | 40 | 40 | 40 |
| GGTGATATGGG | 40 | 40 | 41.67 | 40 |
| GTTTATAAAAT | 40 | 40 | 36.67 | 40 |
| TTTTATAGGTT | 40 | 40 | 40 | 40 |
| ATAATAATGGT | 40 | 40 | 40 | 40 |
| GGGGAAATGTG | 40 | 40 | 40 | 40 |
| TTTGTTTATTT | 40 | 40 | 40 | 40 |
| Average: | 39.87 | 40.13 | 41.80 | 39.87 |
| | Backward/Complement | | | |
| TTGTTGAATAA | 40 | 40 | 48.89 | 48.89 |
| GAGTTGAAGGA | 40 | 40 | 40 | 40 |
| GGATGATGGGG | 40 | 40 | 42.22 | 44.44 |
| AGGGGTAAAAG | 40 | 40 | 44.44 | 46.67 |
| AGGATGAAGGT | 40 | 40 | 42.22 | 40 |
| GAGGAAGGTGA | 40 | 40 | 42.22 | 40 |
| AAAAGAGTTTG | 40 | 40 | 46.67 | 44.44 |
| GGTGATATGGG | 40 | 40 | 35.56 | 40 |
| GTTTATAAAAT | 40 | 40 | 26.67 | 33.33 |
| TTTTATAGGTT | 40 | 40 | 40 | 26.67 |
| ATAATAATGGT | 40 | 40 | 33.33 | 35.56 |
| GGGGAAATGTG | 40 | 40 | 55.56 | 60 |
| TTTGTTTATTT | 40 | 40 | 40 | 48.89 |
| Average: | 40.00 | 40.00 | 41.37 | 42.22 |

Table 3.2: Classification of single cytosine $motifs$

| | Forward/Template | | | |
|---|---|---|---|---|
| Motif Sequence | $\{\Delta\mu\}$ | $\{\Delta\mu, \Delta\sigma\}$ | $\{\Delta\mu, P\}$ | $\{\Delta\mu, \Delta\sigma, P\}$ |
| TTTTGCTGAGT | 78.33 | 86.67 | 83.33 | 85 |
| AAGTTCAAAAT | 40 | 51.67 | 46.67 | 35 |
| AGATGCAGGGG | 70 | 73.33 | 83.33 | 76.67 |
| AAGGGCTGGAT | 66.67 | 75 | 68.33 | 61.67 |
| ATTTGCTGAGG | 65 | 78.33 | 58.33 | 73.33 |
| TGGGGCAAATG | 68.33 | 75 | 68.33 | 83.33 |
| GGAATCAAATT | 40 | 40 | 40 | 40 |
| GTGGACAGGAA | 76.67 | 68.33 | 71.67 | 75 |
| AAATTCTTGAA | 46.67 | 56.67 | 56.67 | 58.33 |
| GAAGACGAAAG | 81.67 | 66.67 | 85 | 76.67 |
| AATGTCATGAT | 53.33 | 61.67 | 68.33 | 73.33 |
| GGTTTCTTAGA | 45 | 43.33 | 58.33 | 51.67 |
| TTTTTCTAAAT | 43.33 | 40 | 60 | 60 |
| Average: | 59.62 | 62.82 | 65.26 | 65.38 |
| | Backward/Complement | | | |
| TTTTGCTGAGT | 53.33 | 51.67 | 64.44 | 66.67 |
| AAGTTCAAAAT | 41.67 | 43.33 | 62.22 | 60 |
| AGATGCAGGGG | 61.67 | 63.33 | 84.44 | 68.89 |
| AAGGGCTGGAT | 63.33 | 63.33 | 73.33 | 73.33 |
| ATTTGCTGAGG | 51.67 | 70 | 75.56 | 73.33 |
| TGGGGCAAATG | 71.67 | 68.33 | 75.56 | 66.67 |
| GGAATCAAATT | 40 | 41.67 | 40 | 57.78 |
| GTGGACAGGAA | 40 | 40 | 46.67 | 40 |
| AAATTCTTGAA | 35 | 40 | 62.22 | 77.78 |
| GAAGACGAAAG | 73.33 | 55 | 80 | 75.56 |
| AATGTCATGAT | 40 | 40 | 57.78 | 62.22 |
| GGTTTCTTAGA | 45 | 46.67 | 68.89 | 66.67 |
| TTTTTCTAAAT | 46.67 | 48.33 | 68.89 | 64.44 |
| Average: | 51.03 | 51.67 | 66.15 | 65.64 |

Table 3.3: Comparison of different HDP topologies and one non-HDP model for three-way classification of cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine. MLE is the maximum likelihood estimate of a normal distribution. The SingleLevel HDP is an HDP model with no subgroupings of 6-mers, Multiset, Composition, MiddleNts, and GroupMultiset HDPs are three-level HDP models described in the results

| Model | Mean Accuracy | Median Accuracy |
|---|---|---|
| MLE | 62% / 50% | 58% / 47% |
| Single-Level | 74% / 66% | 79% / 72% |
| Multiset | 74% / 67% | 80% / 76% |
| Composition | 73% / 66% | 78% / 71% |
| Middle-nucleotides | 71% / 63% | 76% / 69% |
| Group Multiset | 73% / 65% | 78% / 71% |

accuracy that would be expected by chance. They are also significantly better than the results of the HMM with the emissions modeled by normal distributions, which achieved mean and median accuracy of 58% and 62%, respectively, for the template reads and 47% and 50% for the complement reads Figure 3.5A.

The classification accuracy varied substantially among different cytosines on the DNA substrate (Figure 3.5B). The best-performing three-way model classified different cytosines at accuracies ranging from 16% to 95% with median accuracy of 76% for template reads and 70% for the complement reads (Figure 3.5C). Calling of sites as unmodified cytosine is the most common error, with 5-mC being the most commonly miscalled variant 3.5C. It is likely that some of the difficulty in classifying certain sites results from 6-mer ionic current distributions that vary only slightly between the methylation states. Therefore, we compared the mean pairwise Hellinger distance between the ionic current distributions of the methylation states of the 6-mers overlapping a site and the sites classification accuracy. The Pearson correlation was 0.52 ($p = 6.6E\text{-}33$) on the template strand and 0.36 ($p = 9.0E\text{-}15$) on the complement strand Figure 3.5D. This

suggests that there is indeed a relationship between the similarity of the distributions and methylation calling accuracy.

### 3.5.1 The hierarchical Dirichlet process more realistically models ionic current distributions

Figure 3.5.1 compares the current signal distributions of three representative 6-mers from the HDP, the maximum likelihood estimate (MLE) normal distribution, and a kernel density estimate. Qualitatively, compared to MLE, the HDP posterior densities reflect the nuance of the 6-mer distributions more realistically. As a nonparametric method, the HDP can approximate any empirical distribution with sufficient data. The statistical shrinkage between the distribution estimates also tends to smooth away small-scale irregularities that can be observed in the kernel density estimate.

## 3.6 Mapping 6-methyladenine and 5-methylcytosine in genomic *E. coli* DNA

The HMM-HDP model can map modifications to multiple bases on a single substrate within a single sequencing run. As a demonstration, we sequenced pUC19 plasmid DNA grown in E. coli containing both dam and dcm methyltransferases. For this experiment we used the newer R9 sequencing protocol. These genomic DNA substrates are completely methylated at C**C**(A/T)GG and G**A**TC, respectively (the bold character indicating the methylated residue) [44]. We sequenced the methylated plas-

mid and an unmethylated PCR amplicon in the same flow cell (Section 3.10.2). The pUC19 DNA sequence is 2,686 base pairs long. It contains 30 adenine residues in GATC motifs and 10 cytosines at C**C**(A/T)GG motifs. The motifs are palindromic, so they each contain two potentially modified residues: one on each strand. The reads covered the entire length of the substrate.

The model was more accurate for cytosine methylation than adenine methylation. The mean per-read accuracy of calling cytosine variants 79% and 72% on the template and complement strands, respectively. The accuracy for calling adenine variants on the two strands was 70% and 58%. The empirical current distributions for cytosine showed a more pronounced difference between methylation states than did adenine (Figure 3.6-top). This probably contributes to the lower accuracy on adenine. To assess how robust the model is to variations in data quality, we explored the relationship between the ungapped alignment score and methylation calling accuracy. The ungapped alignment score reflects many potential sources of variation in read quality. We randomly sampled 40X read coverage and called methylation variants using the posterior probabilities produced by the model. The variation in per-read accuracy was correlated with the ungapped alignment score (Figure 3.6-bottom).

Consensus classification results are shown in Supplementary Table 2A with various thresholds on ungapped alignment score and posterior probability. The best parameters classified 96% and 91% of the residues correctly for cytosine and adenine residues, respectively. The plasmid sequence only contains five 5-mC motifs. These may not capture enough of variation in sequence context to accurately evaluate the model.

To more rigorously evaluate the models ability to map 5-mC in genomic DNA, we mapped 5-mC at C**C**(A/T)GG motifs in native *E. coli* DNA. We evenly divided 1,709 constitutively methylated sites (containing 3,418 potentially methylated cytosines) into a training and testing set. We assume that none of cytosines in the PCR reads are methylated and all of the cytosines in the stationary phase reads are methylated. Based on these labels, the model was able to correctly classify 96% of the cytosines motifs in the test set. These results show that the method described is suitable for mapping DNA methylation in bacterial chromosomal DNA.

## 3.7 Assaying dynamic methylation levels in genomic DNA

Chemical modifications to DNA happen post-replication, and they are often influenced by the state of the cell. One cell state that is known to affect methylation levels in *E. coli* is growth phase. It takes time for methylation to become established on newly synthesized DNA and DNA-binding proteins compete with methyltransferases for occupancy on the chromosome [44]. We exploited this fact to further test our models sensitivity. We sequenced genomic DNA isolated from *E. coli* cultures harvested at three different growth phases: early-exponential (0.4 OD), late-exponential (0.8 OD), and stationary (24 hours).

To assay cytosine methylation we employed the model described in the previous section, except that we trained it on all 3,418 known-methylated cytosines. We trained the model on reads from stationary phase genomic DNA and PCR amplified DNA. To

train the adenine classification model we labeled all adenines at GATC sites in the E. coli genome as methylated in reads from stationary phase cells (Section 3.9.2). In total, we classified 24,100 cytosines at CC(A/T)GG motifs and 38,248 adenines at GATC motifs. The results from both classification experiments are shown in 3.7.

Our model produced cytosine methylation calls consistent with known patterns across growth phases. In the cytosine experiments, we used both template and complement reads and called 23,004 (95.5%), 23,789 (98.7%), and 24,034 (99.8%) of the cytosines as methylated in the early-exponential, late-exponential, and stationary growth phases respectively. These results are consistent with previous studies that showed increasing levels of cytosine methylation from early-exponential phase growth through stationary phase growth [23]. We evaluated the model by classifying the 3,418 known cytosines using PCR and stationary phase reads that were held out during the training of the model. The accuracy and precision were 96% and 92%, respectively.

The adenine methylation calls were also consistent with known biology. Our adenine classification experiments on pUC19 plasmid DNA showed that using only template reads gave the highest accuracy adenine methylation variant calls (Table 2A). Therefore, we only used template reads to assay for genomic adenine methylation levels. The classifier called 33,930 (89%), 34,884 (91%), and 31,901 (83%) of the adenines as methylated in the early-exponential, late-exponential, and stationary growth phases respectively. Transcriptional levels of dam have been shown to correlate with growth-rate, reaching a maximum during exponential phase growth, followed by a decrease during stationary phase growth 5,23. Our results are consistent with this pattern. We did not

have a mapping of 6-mA in the *E. coli* genome to test the model with. Instead, to evaluate the accuracy of the model we classified adenine variants on the pUC19 plasmid using the iterative procedure described previously. The model trained on approximate labels had an estimated accuracy and precision of 87(+/-)3% and 84(+/-)4%, respectively. However, the pUC19 sequence does not contain all of the GATC contexts in the *E. coli* genome, so this measure of accuracy may not fully generalize.

## 3.8 Mapping 5-methyl cytosine in human genomic DNA

We employed the cloud-based `SignalAlign` algorithm (Chapter 2 and this chapter), to map 5-methyl cytosine at CpG dinucleotides on chromosome 20 of the GRCh38 reference. Compared to the analysis in *E. coli* the dataset required to process a single chromosome was roughly 10 times larger, and thus required significantly more computational resources, explained in Section 2.7. We compared the methylation calls produced by `SignalAlign` to published bisulfite data. Overall we observed good concordance between the two call sets producing an r-value of 0.779. It should be noted that `SignalAlign` produces a *probability* of methylation instead of a hard call, and this likely negatively effects the r-value. We also produced a receiver operating characteristic curve the probability of methylation to the number bisulfite reads reporting methylation and produced an area under the curve of 0.9 (Figure 3.8). To show the general agreement in the trends, we plotted the probability of methylation along with the confidently called CpGs (ones where the calls were either all methyl or all not methyl cytosine) for

a stretch of chromosome 20 Figure 3.8.

## 3.9 Data selection and partitioning for model training

### 3.9.1 Dividing *E. coli* methylation motifs into training and test groups

From the results of bisulfite sequencing performed on stationary phase *E. coli* K12 MG1655 we parsed 1,709 motifs, **A**, with the sequence `CCWGG` where the innermost cytosine is methylated [23] (`W` refers to either `A` or `T`). Analysis of the genome showed that 456 different 6-mers occur at the motifs centering around the second cytosine, let this set be denoted as $\mathcal{K}$. A training group of motifs, **T**, was generated by randomly drawing motifs from **A** until all 6-mers in $\mathcal{K}$ were observed. Additional motifs from **A** were added at random until **T** contained roughly half of the total motifs. The remaining motifs were assigned to the test group, **R**, such that $\mathbf{T} \cap \mathbf{R} = \emptyset$. The same groupings were used in experiments with 5-mers. The HDP-HMM was trained on alignments generated with pcrDNA reads supplemented with events from gDNA reads that aligned to the high-confidence methylated sites from the training group. We used the trained model to classify the methylation status of cytosines in the test group motifs from the held out portion of reads from the pcrDNA and gDNA sequencing runs.

### 3.9.2 Adenine classification with approximate labels

The HMM-HDP uses a supervised learning method where k-mer/event pairs are fed to an MCMC algorithm (section 2.3). Unlike the case with 5-mC, we did not

have a high-confidence mapping of 6-mA for *E. coli* K12 and it has been shown that adenines at GATC contexts are not entirely methylated during stationary phase growth [37, 11, 43]. To train the model to learn the new distributions attributed to 6-mA we labeled all of the adenines at GATC motifs (two per motif) as 6-mA in reads from stationary phase genomic DNA. We used events from PCR-generated reads to learn the canonical base k-mers. The improperly labeled un-methylated adenines in the training data likely explain the relatively high rate that adenines care called 6-mA in the PCR data. The effect is not overwhelming, however, as changes in the genomic methylation level is still detectable. In general, this shows that the model can be used to assay methylation when perfectly labeled training data is not available (see 2.6.6 for further discussion).

## 3.10 Sequencing materials and methods

### 3.10.1 MinION sequencing

The sequencing runs on synthetic oligonucleotides were performed in late 2015 using R7.3 chemistry (SQK-MAP006 sequencing kits). The R7.3 MinION sequencing protocol records ionic current at 3kHz and modeled event as corresponding to 6-mers. The pUC19 plasmid DNA, E. coli native and whole-genome-amplified DNA were sequenced using R9 chemistry (EXP-NSK007 sequencing kits). The R9 version uses a different pore and increased sequencing speed. In this version of the protocol, the MinION samples ionic current at 4kHz and the events are modeled as 5-mers. We initially

used a 6-mer lookup table for the R9 pore provided by ONT, then estimated our own 5-mer model from a collection of reads (Section 3.1).

### 3.10.2 Sequencing controlled synthetic DNA substrates containing C, 5-mC, or 5-hmC

We used 897 bp synthetic DNA oligonucleotides from ZYMO Research (Catalog # D5405) that contain entirely C, 5-mC, or 5-hmC bases. Apart from the cytosines, the oligonucleotides have identical sequences. We performed sequencing experiments using R7.3 chemistry (SQK-MAP006 sequencing kits) with four MinION flow cells: one for each of the three substrates, and one where all the substrates were with barcoded with uniquely identifying sequences (EXP-NBD001 barcoding kit) and run together on one flow cell. The runs where the strands were sequenced individually produced 68,920, 27,073, and 70,641, reads for the C, 5-mC, and 5-hmC strands, respectively. The run where the strands were barcoded and sequenced together produced 6,966, 294, and 467 reads for the C, 5-mC, and 5-hmC strands, respectively. All models were trained on the reads where the strands were run in separate flow cells. The bar-coded reads served as our test dataset. This experimental design maximized the amount of training data while controlling for batch effects between MinION runs. Sequence data were processed using Metrichor (versions 1.15.0 and 1.19.0), and only "pass" 2D reads that covered the full length of the reference sequence were used for downstream analysis.

### 3.10.3 Preparation of DNA control substrates containing 6-mA and 5-mC

We purchased pUC19 vector DNA from New England Biolabs (NEB cat. number N3041S). This DNA is isolated from E. coli strain ER2272 that contains genes methyltransferase (MTase) genes *dam* and *dcm*. The dam MTase methylates the adenine in GATC sequence contexts and the *dcm* MTase methylates the inner cytosine at C**C**(A/T)GG sequence contexts. We linearised the plasmid by restriction digest at a unique SspI (NEB cat. number R0132S) restriction site. The linearised plasmid was purified by excising the band from an agarose gel following electrophoresis. The DNA was eluted from the gel using the Wizard SV kit (Promega) as per manufacturer's instructions. To generate an unmethylated substrate, we PCR amplified the plasmid with primers around the SspI restriction site (forward: 5 ATT ATT GAA GCA TTT ATC AGG GTT ATT GTC, reverse: 5 ATT GAA AAA GGA AGA GTA TGA GTA TTC AAC) with Q5 high-fidelity polymerase master mix (NEB cat. number M0492S) as per the manufacturer's specifications. The PCR reaction was purified with 0.4X AMPure SPRI beads using standard procedures.

### 3.10.4 Sequencing for pUC19 plasmid DNA

The purified PCR-amplified and linear pUC19 DNA were individually barcoded (EXP-NBD002 barcoding kit). Roughly equimolar amounts of the barcoded material was combined and sequenced on the MinION using R9 chemistry (NSK-007 sequencing kit). The sequencing run produced 27,293 and 17,220 pass 2D reads in the

PCR-identified and native-identified categories, respectively.

### 3.10.5 Sequencing for genomic and amplified E. coli DNA

We performed one sequencing run using standard procedures on genomic DNA (gDNA) from E. coli strain K-12 MG1655 and another run on DNA that was PCR-amplified (pcrDNA) using a whole-genome amplification kit (Qiagen REPLI-g). Both runs were done independently using R9 chemistry (EXP-NSK007 sequencing kits). The gDNA run produced 18177 pass 2D reads (132 Mb) with an average read length of 7.3 kb. The pcrDNA run produced 61408 pass 2D reads (387 Mb) with an average read length of 6.3 kb. The reads were shuffled and evenly divided into two groups, one was used for training the model and the other for classification experiments.

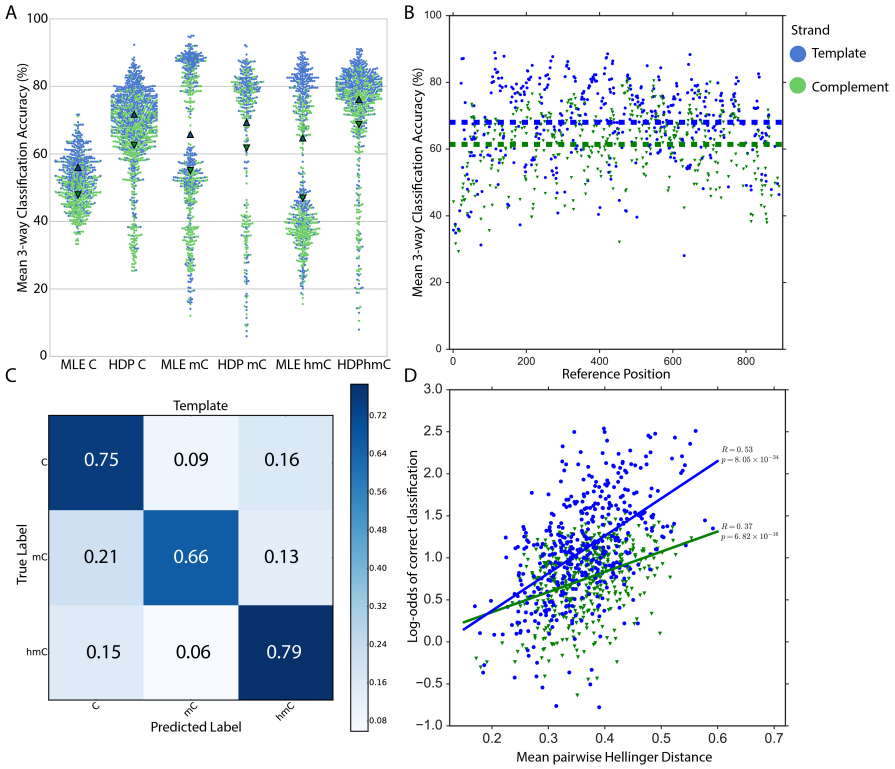


Figure 3.1: Cytosine methylation variant calling accuracy results on synthetic oligonucleotides. Results are from classification of 6,966, 294, and 467, C, 5-mC, and 5-hmC strands respectively that were barcoded and sequenced in the same MinION flow cell. A. Per-read accuracy distribution is shown for the maximum-likelihood estimate (MLE) normal distributions and the Multiset HDP model. The triangles represent the mean of the distribution. B. Average three-way classification accuracy for all sites on the substrate. Dotted lines represent the mean across all sites for template (blue) and complement reads (green). C. Confusion matrix showing HMM-HDP three-way cytosine classification performance on template reads of synthetic oligonucleotides. D. Scatter plot showing the correlation between the log-odds of correct classification and the mean pairwise Hellinger distance between the methylation statuses of the 6-mer distributions overlapping a cytosine

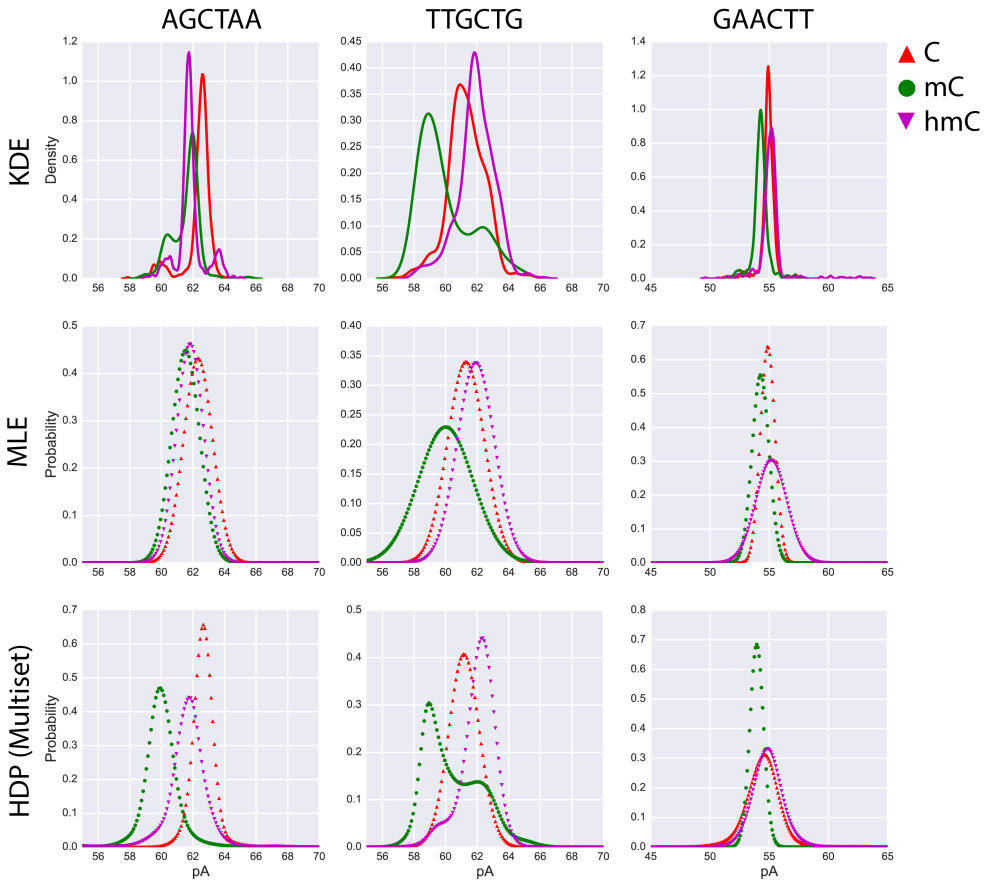


Figure 3.2: Probability distributions for three representative 6-mers by multiple methods. The first row shows the kernel density estimate (KDE) based on the preliminary alignments described in the text. The middle row shows maximum likelihood estimated (MLE) normal distribution probability density functions. The bottom row shows probability density functions from the Multiset hierarchical Dirichlet process (HDP). All data shown are from template reads.

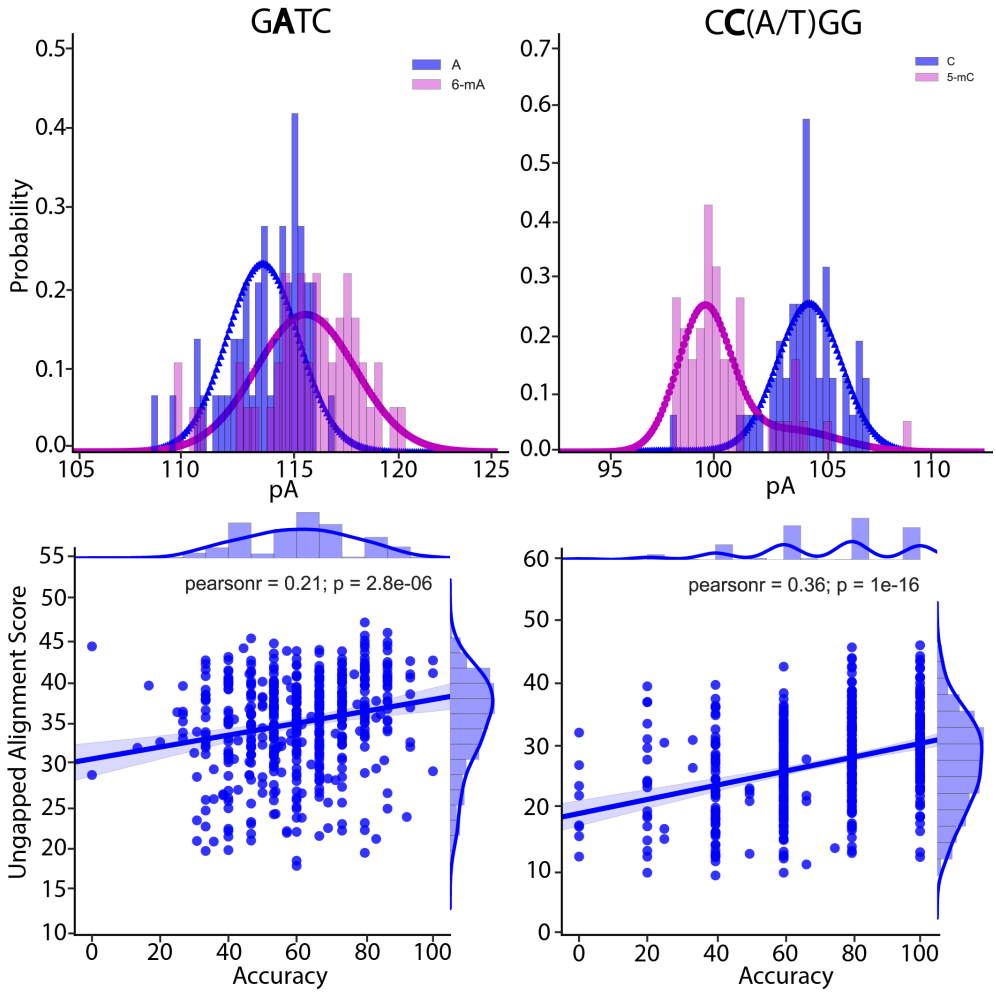


Figure 3.3: Observed and learned ionic current distributions and read accuracy correlation with ungapped alignment score for 6-mA in GATC motifs (left) and 5-mC in CC(A/T)GG motifs (right). Top: Comparison of the influence of 6-mA and 5-mC on ionic current levels for representative 5-mers. The empirical ionic current levels from 100 aligned events are shown as a normalized histogram and the HDP-learned probability densities were are shown as curves. The HDP density was sampled on 900 point grid from 50 to 140 pA. Bottom. Correlation between ungapped alignment score (see Methods) and per-read accuracy for 500 randomly sampled template reads

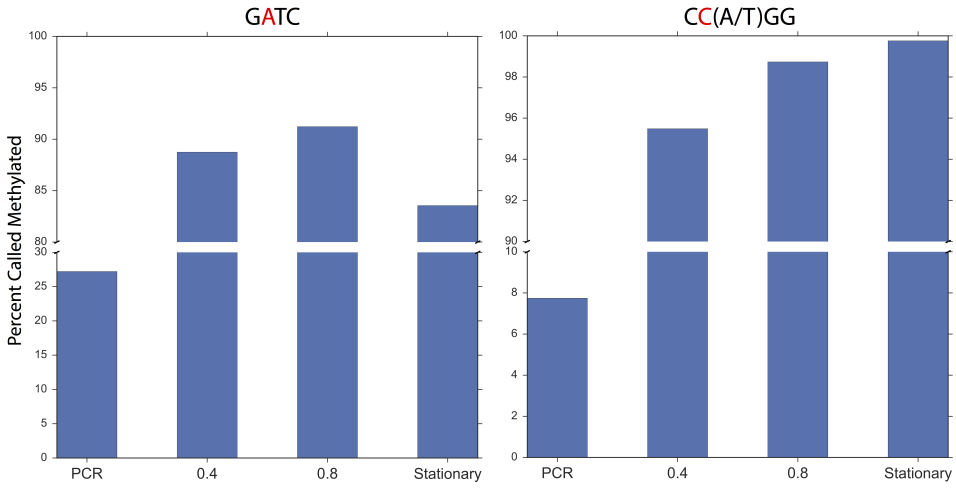


Figure 3.4: Changes in genome-wide cytosine methylation at different stages of culture growth. Bar height represents the percentage of residues that were called as methylated. Axes are broken to accentuate differences between the growth phases.

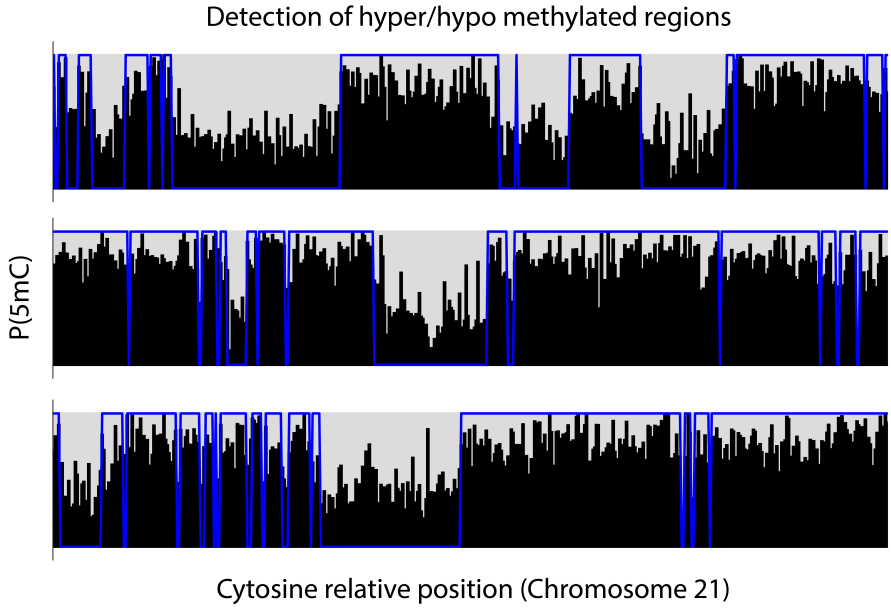


Figure 3.5: Qualitative concordance between `SignalAlign` methylation probabilities (black bars) for 1,500 human CpG dinucleotides on chromosome 20, blue line shows the "true" bisulfite calls.

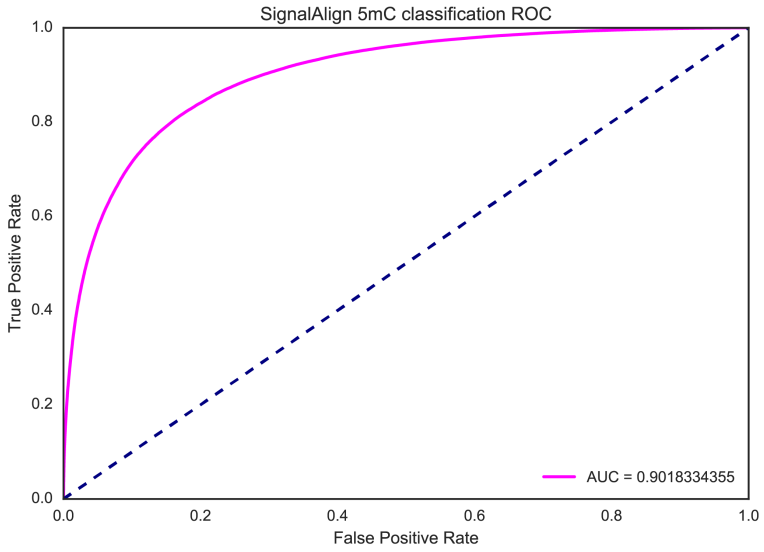


Figure 3.6: Receiver operating characteristic curve showing the classification performance of `SignalAlign` on chromosome 20.

# Chapter 4

# Rereading DNA with helicase *Hel308*

## 4.1 Introduction

Due to the heterogeneity of the epigenome high-thoughput and sensitive sequencing techniques must be devised that are sensitive to multiple cytosine modifications simultaneously [48]. Existing technologies such as TET-assisted and oxidative bisulfite sequencing, where only unmodified cytosines are chemically modified to uracil allows for separate detection of methyl and hydroxymethylcytosine after amplification by comparing treated and untreated samples can only detect one modification at a time [5, 53]. Single-molecule real time (SMRT) sequencing harnesses the kinetics of polymerase dNTP incorporation at modified bases on the template strand and relies on rereading the insert sequence for increased accuracy [17].

The primary signal of nanopore sequencers is the sequence specific partial blockage of ionic current by DNA. Using a DNA polymerase to control strand transloca-

tion, nanopore based platforms have been shown to discern between C, methyl-cytosine (mC), hydroxymethyl-cytosine (hmC), formyl-cytosine, and carboxy-cytosine at accuracy rates ranging from 87-98% based on a single read of a DNA molecule [46, 52]. Commercial nanopore based sequencers read DNA molecules in the range of 0.5-48 kb in length. An advantage of single molecule sequencing technologies is the opportunity to characterize sub-populations within a sample. Thus a goal for these sequencers is to increase accuracy of mapping mC and hmC bases on an individual molecules. One way to increase calling accuracy is to reread a single molecule multiple times. If errors are independent and randomly distributed, then multiple reads can be used to gain more information of a target substrate.

This section describes a nanopore system where a Hel308 helicase from Methanococcoides burtonii (Hel308Mbu, referred to as Hel308) pulls single-stranded DNA (ssDNA) against an applied voltage through a mutated Mycobacterium smegmatis porin A (M2MspA). Synthetic DNA strands containing a C, mC, or hmC embedded within a CC*GG context (C* denoting the modified base) and an independent label sequence were used to determine if rereading can increase accuracy of classification of modifications. A hidden Markov model (HMM) was used to model the movement of Hel308 on the DNA strand and multiple analysis algorithms were implemented to determine the accuracy of single and multiple reads.

## 4.2 A method to reread DNA

Nanopore sequencing reads native DNA by directly touching the individual bases in the strand. We engineered a system that can pass a single molecule of DNA through the nanopore multiple times, this strategy we call the *Break Away* system. A schematic of the design can be seen in Figure 4.2. Hel308 binds to a 5-nucleotide single-stranded loading site. To prevent helicase activity in bulk, a G-quadruplex forming sequence is positioned after the loading site. The 5 end of the strand contains a run of abasic residues that causes Hel308 to dissociate. A complement strand that bears a 3 cholesterol aids in capturing the substrate by concentrating the substrate at the membrane. Once the strand is captured and threaded through the pore the complement is removed and the force from the applied voltage is sufficient to unfold the G-quadruplex [35]. The resulting ssDNA allows Hel308 to translocate on the strand, however because Hel308 cannot pass through the MspA pore it pulls the strand through MspA against the applied voltage. Based on crystal structures Hel308 occludes about 11 nucleotides [6].

A representative trace from an event with 4 rereads is show in Figure 4.2. Briefly, the strand is captured and the first read starts immediately (Read i). Next, the abasic block passes through the pore sensor resulting in high ( 82 pA) current (red bars). Hel308 arrives at the abasic block shortly after it passes through though the pore and the first enzyme dissociates. This causes the DNA to fall back to just before the abasic block (about 11 nucleotides) where another Hel308 is positioned. The process

repeats itself two additional times until all enzymes are removed from the strand. In the case of the first read there were 3 Hel308 on the strand. Once all of the enzymes are removed, the strand falls back to the starting position and the process repeats itself affording the second, third, and forth read. Occasionally the read will fail to restart either from the G-quadruplex not refolding or because too many Hel308 are loaded on the strand, in this case we observe reads that do not start from the beginning but rather somewhere in the middle of the strand (green bars).

With the G-quadruplex unfolded plus the single-stranded loading site there are 20 nucleotides available, allowing for multiple Hel308 enzymes to bind. This is advantageous because these enzymes have been shown to act as functional oligomers [47]. In the case where multiple Hel308 have loaded on the strand, a 10-12 nucleotide backslip is observed instead of a reset to the start of the read Figure **??** (red bars). We typically observe 2-3 enzymes bound per-read. Once all of the active Hel308 are removed, the strand remains captured and is pulled by the voltage to the G-quadruplex sequence and the strand is reread.

## 4.3 Mapping ionic current segments to 4-nucleotide words

To determine the segments corresponding to the modified cytosine in the context we mapped each segment in the read to its corresponding nucleotides in the sequence. It has been shown that 4 nucleotides (4-mers) are responsible for the segment current level when using MspA [26, 52]. Within a read we observed Hel308 can occa-

sionally backslip 1-5 nucleotides (discussed in the next section), a property that can make mapping current segments to 4-mers difficult in repeating regions. Two strategies were implemented to unambiguously map current segments to nucleotide 4-mers. First, asymmetric bookend homopolymer sequences were added to the 5 and 3 end of the sequence of interest (Figure 4.3). The bookend sequences allow for the accurate mapping of the first and last non-bookend 4-mer. Second, we used a sequence where each 4-mer has unique neighbors within the sequence context. This means that even though a segment may occur more than once in the read it can be mapped to its 4-mer based on its immediately following segment. One 4-mer, TCAT, occurs 3 times within the sequence. The TCAT segments were used as anchor points allowing for mapping of the intermediate segments (4.3).

The sequence employed contains a context and label region (Figure 4.2). The context region consists of the CpG dinucleotide where the cytosine can be C, mC, or hmC and the neighboring 5 and 3 nucleotides. To emphasize the advantage of rereading we chose to use the CCGG context because it had low single-read accuracy in previous studies [52]. The label sequences were chosen to be GG, AA, and TT for C, mC, and hmC respectively.

## 4.4 Modeling and classification of reads using modular HMM

Hidden Markov Models (HMMs) have been used in many computational applications including statistical modeling, database searching, and multiple sequence alignment of protein families and domains [7]. Similar to other SF2 helicases we observed Hel308 backslip one or more nucleotides causing a repeat of segments to be observed [8]. Hel308 can also dissociate from the substrate mid-read, and if another helicase happens to be bound directly behind it (a common occurrence), the DNA will fall back the length of one helicase occlusion site (10 nts, 6 segments). To model this behavior, a standard profile HMM was used as a reference point with an added modular structure and a backslip pathway that models both spontaneous backslip and mid-read Hel308 dissociations (Figure 4.4 bottom). To more accurately represent the biophysical similarity between backslips and dissociations probabilities were selected such that backslips of 1-5 nucleotides carry an exponential decay in transition probability and yet the probability of a backslip of 1 segment is equivalent to a backslip of 6 segments (Figure 4.4 bottom). The HMM was designed with forks for the 3 modifications and their corresponding labels. Taking these differences into account led to the complete model seen in Figure 4.4-top, that allows the different substrates to align to their respective track in the fork.

Evaluation of multiple algorithms for combining single-molecule rereads For events that contain rereads, we computed the accuracy based on 3 methods, Best,

independent consensus (IC), and standard posterior decoding (PD). The Best method calculates the accuracy based on the calls from the Chunk with the highest CScore. The IC method calculates the error of a read by the product of the probabilities of error of each Chunk, following the equation: . A full discussion of the IC method can be found in the supplement. The PD method calculates the accuracy by computing the sum of the probability over all paths through the HMM.

Of the three multi-read algorithms IC produces the highest accuracy for reads where the context CScore is greater than 0.65 (Figure **??**A). Using the IC algorithm events in this range have 11-18% higher accuracy if they contain multiple Chunks. Chunks with intermediate CScores (0.65 ¿ CScore ¿ 0.35) do not benefit from multi-reads over single reads using IC, and when the CScore is less than 0.35 single reads perform better than IC. For a distribution of the number of reads in the events see Figure 4.4B. We can reconcile this decrease by using the Best algorithm for Chunks with CScore below 0.35 (Figure **??**C). A confusion matrix was constructed to show how each context was being called. C is miscalled as mC and hmC equally, whereas both mC and hmC have a bias of being called as cytosine.

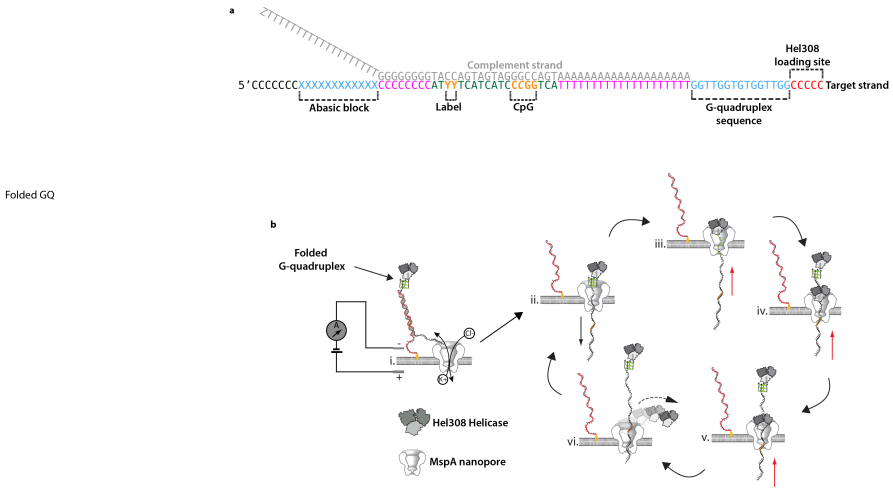


Figure 4.1: A. Schematic of typical nanopore setup. A single MspA porin is inserted into an artificial lipid membrane suspended between a teflon aperture. The membrane separates two buffered solutions into the cis and trans compartments. A voltage is applied across the membrane and the resulting ionic current is monitored using a patch-clamp amplifier. B. Scheme of the Break-Away reread system: i. The hybrid substrate is attracted to the membrane by a 3 cholesterol-linked tether oligo that is partially complementary to the reread strand. Hel308 helicase in bulk loads onto the 3 end of the reread strand, behind the folded G-quadruplex. ii. The 5 tail of the reread strand threads through MspA, removing the tether strand. Electrophoretic force pulls the reread strand until the G-quadruplex is positioned between Hel308 and the constriction of MspA. iii. The force from the voltage unfolds the G-quadruplex into single-stranded DNA allowing Hel308 to process on the strand in the 3 to 5 direction. iv. Hel308 pulls the reread strand against the voltage allowing the strand to be read by recording the changes in ionic current through MspA. Simultaneously, the G-quadruplex refolds preventing additional Hel308 from processing on the reread strand. v. The block of abasic residues passes through MspA, producing a characteristic high (85 pA) current level. vi. When Hel308 reaches the block of abasic residues, it dissociates. Allowing the reread strand to return to position ii, initiating a reread.

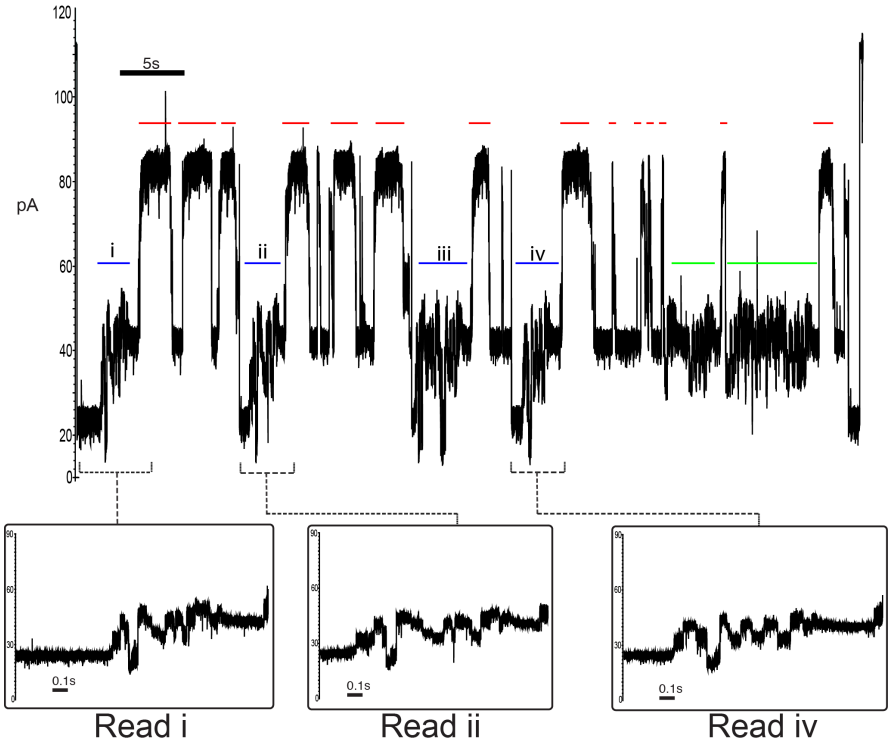


Figure 4.2: Representative current trace with 4 complete rereads of a single DNA molecule. This current trace is from due to the mC context and AA label. High current levels from abasic resides are highlighted by red bars (Fig. 1B v.). Full length reads with bookends are shown by blue blue bars and expanded below. Unsuccessful restarts, where the read does not start from the beginning (pol-dT bookend) are shown by green bars.

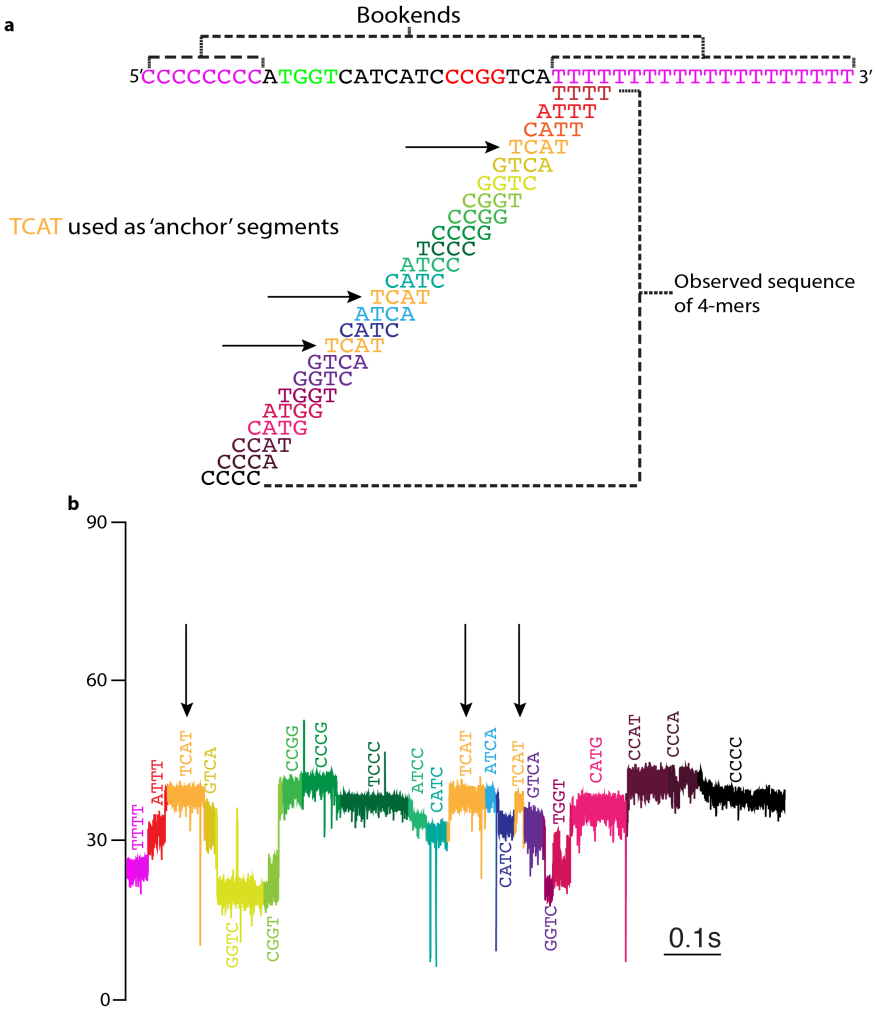


Figure 4.3: A. Cartoon representation of entire substrate used in Break-Away system. The 5 poly-dC tail threads through MspA. The abasic block, represented as a is 12 residues long to ensure Hel308 will dissociate. The sequence of interest contains the 5 poly-dC bookend, the CC*GG context sequence, the corresponding label sequence, and the 3 poly-dT bookend. The 3 end of the strand contains a 5-cytosine Hel308 binding sequence and a thrombin binding aptamer sequence that folds into a G-Quadruplex. B. Reread sequence showing mapped 4-mers and bookend sequences. Reoccurring TCAT 4-mers were used as anchor states with consistent 43.5 pA current. The sequence was designed so that each 4-mer has unique neighbors within the sequence (see text). C. Representative current trace with mapped segments colored according to 4-mer sequence as in B. Arrows show TCAT anchor segments.

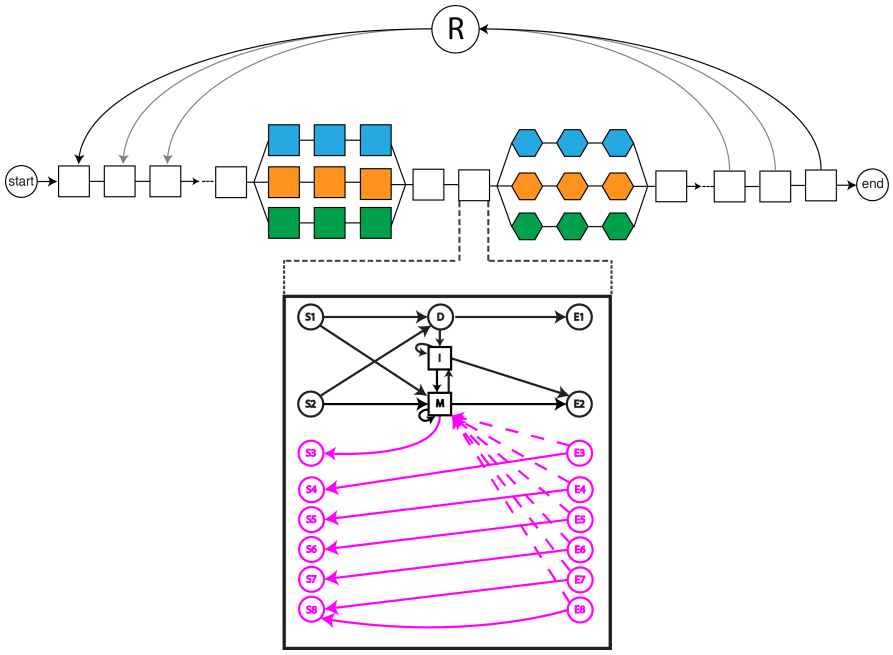


Figure 4.4: A. (Top) Each current level in the mapped ionic traces was mapped as a separate board (below) allowing for transitions to and from match/insert, delete, and back slip paths through the model. The current states corresponding to the context and label were modeled as three separate paths through the model and used for classification. The downstream label (hexagons) were used to confirm or disprove the classification of the context. (Below) A modular board in the HMM representative of a specific segment in a nanopore trace. Circular nodes represent silent states (non-emitting states), D is the delete state (missing segment), I is the insert state (off-pathway segment/noise spikes), M is the match state (aligning to a segment of the same mean), and the red states represent the backslip pathway. B. There are two roughly equally likely possibilities that can explain a back slip observation; one Hel308 can move backwards on the reread strand or one Hel308 can dissociate mid-read and the strand will move backwards to a trailing Hel308.

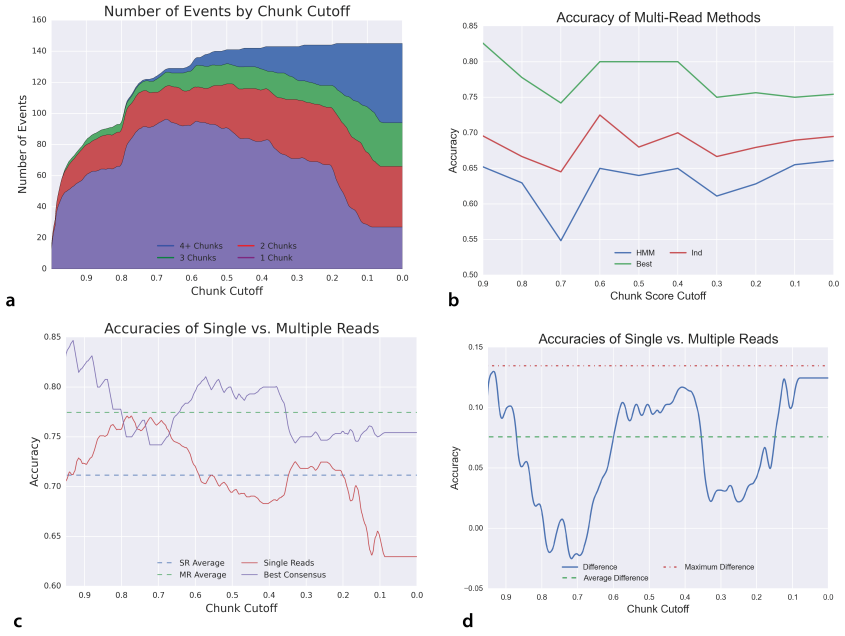


Figure 4.5: A. Plot comparing accuracy of multi-read events to single-read events by CScore. At CScore>0.65 multi-read evens have accuracies 11-18% higher than their single-read counterparts. From 0.65>CScore>0.35 single and multi-read events are roughly the same. In events with CScore<0.35 single reads are more accurate. B. Plot showing the number of events used and the number of Chunks contained in those reads above a given CScore. C. Plot of accuracy by CScore by various methods. Best and IC are described in the text. First, Last, and Random were used as controls for ordering bias. IC is the highest performer at CScore>0.3, after which Best is the highest performer. D. Confusion matrix showing occurrence of miscalls by type.

# Chapter 5

# Conclusion

In this work, I developed new models for detection of DNA base modifications with nanopore sequencing instruments. In Chapter 2 I describe a hybrid statistical model composed of a hidden Markov model and hierarchical Dirichlet process mixture of normal distributions. In Chapter 3 I use the model in combination with the MinIONs ionic current data, I showed how the model can achieve a median three-way cytosine classification read accuracy of 80% on synthetic DNA. I explain that the classification accuracy varies between sequence contexts and is correlated with the impact the methylation has on the ionic current signal. In Section 3.6 I show how I tested the method in a model system by mapping 5-mC and 6-mA bases in genomic E. coli DNA and plasmid DNA. The described model correctly mapped the methylation status of 96% of the cytosines in *E. coli* DNA and 86% of the adenines in pUC19 plasmid DNA with 20X and 40X coverage, respectively. This coverage is lower than is typical in studies with other platforms. To show the utility of the method in a dynamic system, I demonstrated how

the model can detect genome-wide changes in methylation at different growth phases in *E. coli* even with imperfect training data. Lastly in Section 3.8 I show how the model can be scaled to human genome experiments and retain high performance in detecting CpG methylation. Moreover, since no extra sample preparation is necessary, this information is available in any MinION sequencing experiment that uses genomic DNA.

I anticipate numerous applications for this method. For instance, it could be used to physically phase multiple base modifications simultaneously on long reads, allowing for haplotype phasing of differentially methylated blocks in the genome. This method is also applicable as a starting point for additional machine learning algorithms. For example by probabilistically assigning events to methylation and variant status, new base calling neural networks can be trained to detect base modifications *de novo*. In addition, it is straightforward to change the set of base modifications our model detects as long as there is appropriate training data (Section 2.6.6), for example modified RNA bases or DNA damage.

# Bibliography

[1] M Akeson, D Branton, J J Kasianowicz, E Brandin, and D W Deamer. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophysical journal*, 77(6):3227–33, dec 1999.

[2] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, (December), dec 2014.

[3] Mariam Ayub, Steven W Hardwick, Ben F Luisi, and Hagan Bayley. Nanopore-Based Identi fi cation of Individual Nucleotides for Direct RNA Sequencing. *Nano letters*, 13(12):6144–6150, 2013.

[4] F Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv: ...*, pages 1–10, 2012.

[5] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (New York, N.Y.)*, 336(6083):934–7, may 2012.

[6] 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):647–52, jul 2007.

[7] Christopher Bystroff and Anders Krogh. Hidden Markov Models for prediction of protein features. *Methods in molecular biology (Clifton, N.J.)*, 413:173–98, jan 2008.

[8] Wei Cheng, Srikesh G Arunajadai, Jeffrey R Moffitt, Ignacio Tinoco, and Carlos Bustamante. Single-base pair unwinding and asynchronous RNA release by the hepatitis C virus NS3 helicase. *Science (New York, N.Y.)*, 333(6050):1746–9, sep 2011.

[9] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, 30(4):344–348, 2012.

[10] Scott L. Cockroft, John Chu, Manuel Amorin, and M. Reza Ghadiri. A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *Journal of the American Chemical Society*, 130(3):818–820, 2008.

[11] Nadia R Cohen, Christian a Ross, Saloni Jain, Rebecca S Shapiro, Arnaud Gutierrez, Peter Belenky, Hu Li, and James J Collins. A role for the bacterial GATC methylome in antibiotic stress survival. *Nature Genetics*, (February), 2016.

[12] Joseph M Dahl, Ai H Mai, Gerald M Cherf, Nahid N Jetha, Daniel R Garalde, Andre Marziali, Mark Akeson, Hongyun Wang, and Kate R Lieberman. Direct observation of translocation in individual DNA polymerase complexes. *The Journal of biological chemistry*, 287(16):13407–21, apr 2012.

[13] Joseph M Dahl, Hongyun Wang, Jose M Lazaro, Margarita Salas, and Kate R Lieberman. Dynamics of Translocation and Substrate Binding in Individual Complexes Formed with Active Site Mutants of {Phi}29 DNA Polymerase. *The Journal of biological chemistry*, jan 2014.

[14] Iwijn De Vlaminck and Cees Dekker. Recent Advances in Magnetic Tweezers. *Annual Review of Biophysics*, 41(1):453–472, 2012.

[15] Richard Durbin, Sean Eddy, Anders Krogh, and Biological Sequence Analysis. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.

[16] Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10), 2011.

[17] Benjamin a Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson a Clark, Jonas Korlach, and Stephen W Turner. Direct detection

of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461–5, jun 2010.

[18] Carl W. Fuller, Shiv Kumar, Mintu Porel, Minchen Chien, Arek Bibillo, P. Benjamin Stranges, Michael Dorwart, Chuanjuan Tao, Zengmin Li, Wenjing Guo, Shundi Shi, Daniel Korenblum, Andrew Trans, Anne Aguirre, Edward Liu, Eric T. Harada, James Pollard, Ashwini Bhat, Cynthia Cech, Alexander Yang, Cleoma Arnold, Mirkó Palla, Jennifer Hovis, Roger Chen, Irina Morozova, Sergey Kalachikov, James J. Russo, John J. Kasianowicz, Randy Davis, Stefan Roever, George M. Church, and Jingyue Ju. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *Proceedings of the National Academy of Sciences*, 113(19):5233–5238, 2016.

[19] Brett Gyarfas, Felix Olasagasti, Seico Benner, Daniel Garalde, Kate R. Lieberman, and Mark Akeson. Mapping the position of DNA polymerase-bound DNA templates in a nanopore at 5 ?? resolution. *ACS Nano*, 3(6):1457–1466, 2009.

[20] Breton Hornblower, Amy Coombs, Richard D Whitaker, Anatoly Kolomeisky, Stephen J Picone, Amit Meller, and Mark Akeson. Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods*, 4(4):2006–2008, 2007.

[21] Nicholas Hurt, Hongyun Wang, Mark Akeson, and Kate R. Lieberman. Specific nucleotide binding and rebinding to individual DNA polymerase complexes captured on a Nanopore. *Journal of the American Chemical Society*, 131(10):3772–3778, 2009.

[22] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, (February), feb 2015.

[23] Christina Kahramanoglou, Ana I. Prieto, Supriya Khedkar, Bettina Haase, Ankur Gupta, Vladimir Benes, Gillian M. Fraser, Nicholas M. Luscombe, and Aswin S.N. Seshasayee. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nature Communications*, 3:886, 2012.

[24] Kevin Karplus. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, 37(SUPPL. 2):492–497, 2009.

[25] 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 of the United States of America*, (24), oct 2013.

[26] 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*, (June):2–7, jun 2014.

[27] Kyung Suk Lee, Hamza Balci, Haifeng Jia, Timothy M Lohman, and Taekjip Ha.

Direct imaging of single UvrD helicase dynamics on long single-stranded DNA. *Nature communications*, 4(May):1878, jan 2013.

[28] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G.Craighead, and W. W. Webb. Zero-Mode Waveguides for Singlr-Molecule Analysis at High Concentrations. *Science*, 299(5607):682–686, 2003.

[29] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95, mar 2010.

[30] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[31] Kate R Lieberman, Joseph M Dahl, and Hongyun Wang. Kinetic mechanism at the branchpoint between the DNA synthesis and editing pathways in individual DNA polymerase complexes. *Journal of the American Chemical Society*, 136(19):7117–31, may 2014.

[32] Shixin Liu, Gheorghe Chistol, and Carlos Bustamante. Mechanical operation and intersubunit coordination of ring-shaped molecular motors: insights from single-molecule studies. *Biophysical journal*, 106(9):1844–58, may 2014.

[33] Nicholas James Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. pages 1–11, feb 2015.

[34] Xi Long, Joseph W Parks, Clive R Bagshaw, and Michael D Stone. Mechanical

unfolding of human telomere G-quadruplex DNA probed by integrated fluorescence and magnetic tweezers spectroscopy. *Nucleic acids research*, 41(4):2746–55, feb 2013.

[35] Xi Long and Michael D Stone. Kinetic partitioning modulates human telomere DNA G-quadruplex structural polymorphism. *PloS one*, 8(12):e83420, jan 2013.

[36] 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, 2012.

[37] Martin G. Marinus and Josep Casadesus. Roles of DNA adenine methylation in host-pathogen interactions: Mismatch repair, transcriptional regulation, and more. *FEMS Microbiology Reviews*, 33(3):488–503, 2009.

[38] Jeffrey R. Moffitt, Yann R. Chemla, Steven B. Smith, and Carlos Bustamante. Recent Advances in Optical Tweezers. *Annual Review of Biochemistry*, 77(1):205–228, 2008.

[39] KC Keir C Neuman and Attila Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods*, 5(6):491–505, 2008.

[40] Benedict Paten, Javier Herrero, Kathryn Beal, and Ewan Birney. Sequence pro-

gressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25(3):295–301, 2009.

[41] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H J Baum, Beate Becker-ziaja, Jan Peter Boettcher, Mar Cabeza-cabrerizo, Eeva Kuisma, Christopher H Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, and Janine Michel. Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[42] Arthur C Rand, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten. Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*, 14(4):411–413, 2017.

[43] Lene Juel Rasmussen, M.G. Marinus, and Anders Lobner-Olesen. Novel growth rate control of dam gene expression in Escherichia coii. *Molecular Microbiology*, 12(4):631–638, 1994.

[44] María A. Sánchez-Romero, Ignacio Cota, and Josep Casadesús. DNA methylation in bacteria: From the methyl group to the methylome. *Current Opinion in Microbiology*, 25:9–16, 2015.

[45] J. Schreiber and K. Karplus. Analysis of Nanopore Data using Hidden Markov Models. *Bioinformatics*, (February):1–7, feb 2015.

[46] Jacob Schreiber, Zachary L Wescoe, Robin Abu-Shumays, John T Vivian, Baldan-

dorj Baatar, Kevin Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, oct 2013.

[47] Bartek Sikora, Yingfeng Chen, Cheryl F Lichti, Melody K Harrison, Thomas a Jennings, Yong Tang, Alan J Tackett, John B Jordan, Joshua Sakon, Craig E Cameron, and Kevin D Raney. Hepatitis C virus NS3 helicase forms oligomeric structures that exhibit optimal DNA unwinding activity in vitro. *The Journal of biological chemistry*, 283(17):11516–25, apr 2008.

[48] Charles Swanton and Stephan Beck. Epigenetic Noise Fuels Cancer Evolution. *Cancer cell*, 26(6):775–776, dec 2014.

[49] Y W Teh, M I Jordan, Matthew J Beal, and David M Blei. Hierarchical {D}irichlet Process. *Journal of The American Statistical Association*, 101(476):1566–1581, 2006.

[50] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[51] Winston Timp, Jeffrey Comer, and Aleksei Aksimentiev. DNA base-calling from a nanopore using a viterbi algorithm. *Biophysical Journal*, 102(10):L37–L39, 2012.

[52] Zachary L Wescoe, Jacob Schreiber, and Mark Akeson. Nanopores discriminate

among five C5-cytosine variants in DNA. *Journal of the American Chemical Society*, 136(47):16582–7, nov 2014.

[53] Miao Yu, Gary C Hon, Keith E Szulwach, Chun-Xiao Song, Liang Zhang, Audrey Kim, Xuekun Li, Qing Dai, Yin Shen, Beomseok Park, Jung-Hyun Min, Peng Jin, Bing Ren, and Chuan He. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149(6):1368–80, jun 2012.