UNIVERSITY OF CALIFORNIA
SANTA CRUZ

APPLICATION OF LONG-READ SEQUENCING TO MODIFIED NUCLEOTIDES FOR DETECTING CHROMATIN ACCESSIBILITY

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Brandon Saint-John

September 2023

The Dissertation of Brandon Saint-John
is approved:

Professor Angela N. Brooks, Chair

Professor Christopher Vollmers

Professor Hinrich Boeger

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Application of Long-read sequencing to modified nucleotides for detecting chromatin accessibility

by

Brandon Saint-John

Nucleosomes provide an additional layer of gene regulation by regulating chromatin accessibility. We have found that factors involved in regulating nucleosome positioning, such as SMARCA4, are recurrently mutated in lung adenocarcinoma and can correlate with alternative splicing changes. As a result, it is crucial to understand nucleosome positioning across an entire gene body to understand how they interact to cause changes in splicing. The key to being able to understand this is using long-read sequencing methods to understand the positioning of nucleosomes at once. My thesis focuses on developing methods and computational tools to understand nucleosome positioning with long reads. I developed a sequencing approach called Add-seq that uses a small molecule called angelicin to label accessible regions and determine those positions using nanopore sequencing. Based on the analyses for Add-seq, I also developed a toolkit called `cawlr`. This computational pipeline can automate calling nucleosomes on single molecules from any long-read sequencing technology. This work provides new ways of looking at nucleosomes and understanding their positioning in the context of other biological processes.

To family and friends,

that made every moment worth it.

Acknowledgments

I want to thank my cohort, who helped me be comfortable moving to a new state and build valuable friendships I will carry beyond my years at UC Santa Cruz.

I want to thank all my climbing friends, for bringing me to the outdoors and experiencing nature in an exciting way.

I want to thank all the lab members, for being a sounding board and hear out ideas and all the activities that made lab time enjoyable.

I want to thank my committee, whose valuable insight guided my project and understand new ways to think about my science.

Chapter 1

Introduction

1.1 Epigenetics

Epigenetics represents a secondary layer of regulation that cells use to regulate gene expression, RNA splicing, cell differentiation, and more. Epigenetics regulates through DNA methylation, histone modification, and density of nucleosomes [4, 26, 21]. Nucleosomes comprise a histone octamer complex containing the H2A, H2B, H3, and H4 subunits and 147 base pairs of wrapped DNA [28, 33]. The wrapped DNA is in a structural configuration where DNA-binding proteins such as polymerases and transcription factors cannot bind [2]. The binding of nucleosomes to DNA can fluctuate due to transcription or chromatin remodeling complexes [5]. While there is an optimal sequence that nucleosomes can bind, it can bind to various DNA contexts and depends primarily on a structural DNA motif [32].

1.2 Alternative splicing

Alternative splicing is the process by which exons and introns are spliced from nascent RNA transcripts to form mature RNA transcripts that will be translated into protein [25]. Alternative splicing allows a single gene to produce different proteins by including or excluding exons. The ability of the same gene to produce variations of the same protein is vital for allowing for flexibility in cellular response [35]. Because alternative splicing can cause significant changes in the proteins it produces, errors in regulating alternative splicing can cause diseases [48].

1.3 Mapping nucleosome positions

Because nucleosomes regulate alternative splicing depending on how they are positioned, mapping the positions of nucleosomes can be important for understanding how certain regions are regulated. Several methods have been developed for mapping nucleosomes genome-wide with high-throughput sequencing [8, 39, 22]. These methods rely on fragmenting and enriching DNA corresponding to a single nucleosome. This is because most approaches rely on short-reads that provide information for a single nucleosome and will not be able to provide information about how multiple nucleosomes are positioned, Other methods, such as NOME-seq, have utilized methyltransferases to label DNA not bound by nucleosomes [23, 49]. These methods can find multiple nucleosomes on a single strand but are limited by the short-read sequencer read lengths.

1.4 Long-read sequencing

While short-read methods lack the length to fully understand how nucleosomes control alternative splicing, several new sequencing methods have emerged that have made answering this possible. These methods were notable in that they could sequence long stretches of DNA in single molecules at a lower accuracy. One such method was nanopore sequencing [14]. Nanopore sequencing involves a biological pore that is embedded in a lipid bilayer. A single DNA strand is unwound and ratcheted through the pore with the help of a motor protein. As the DNA passes through the pore, it causes changes in the electrical current running across the pore's sensing region. The sensing region covers approximately five nucleotides and changes the current in a sequence-dependent manner. The current signal can then be translated into the corresponding DNA sequence. These reads are as long as 2.2 megabases [37].

Another central long-read sequencing platform is PacBio-sequencing [40]. This method relies on the circularization of DNA strands. This single circularized strand is bound in a channel called a zero-mode waveguide. A polymerase adds nucleotides bound to a dye that fluoresces a base-specific color each time it is incorporated. Because the DNA is circular, the polymerase will pass over the same DNA multiple times, allowing for greater accuracy by taking the consensus of the signal generated. This method can sequence DNA strands that average 15,000 to 20,000 bases long. The multiple passes allow it to achieve higher accuracy than nanopore sequencing, with a short read length.

Both of these methods can detect modified nucleotides. In nanopore sequenc-

ing, shifts in the electrical current signal for kmers containing the modified can be used to determine whether or not a nucleotide is modified. Previous work has shown these methods can detect modifications as small as cytosine methylation [38, 46]. For PacBio sequencing, modified nucleotides will change the kinetics of nucleotide incorporation. This measurement, the interpulse width duration (IPD), can be used to detect the presence of modified nucleotides [15].

Previous methods have used enzymes to modify nucleotides not bound by nucleosomes and map them with long-reads [29, 1, 51, 47, 45]. These methods have used a variety of methyltransferases for labeling cytosines or adenines. These have used both nanopore and PacBio sequencing. Each has a specific computational pipeline that is used for their specific method.

1.5 Thesis work

In this thesis, we aimed to develop a long-read sequencing approach to understand chromatin accessibility. The eventual goal is to apply this to understand lung cancer and alternative splicing in the context of nucleosomes. In the second chapter, we introduce Add-seq, a novel sequencing method using the small molecule angelicin to label accessible chromatin and identify it with nanopore sequencing. I performed initial sequencing experiments validating the approach and sequencing of the controls. I analyzed and plotted kmer data, comparing modification calls with genome features and orthogonal data. Robert Shelansky performed the analysis of skipped kmers. Namrita

Dhillon and Brett Meisnner performed sequencing of the concentration experiments. In the third chapter, we used our experience analyzing modification data for long-read chromatin accessibility experiments to develop a toolkit called `cawlr`. I implemented the entire pipeline and packaging of the tool myself. Parts of the dynamic programming code were directly adapted from the NP-SMLR paper [51]. We then discuss the future of the field and the applications we plan to use from our results.

Chapter 2

Determining the alternative splicing changes in lung adenocarcinoma from SMARCA4 mutations

2.1 Abstract

Alternative splicing is the process of creating a large diversity of proteins from a small set of genes and is regulated by the positioning of nucleosomes across genes. SMARCA4 encodes a subunit of the SWI/SNF complex responsible for nucleosome remodeling and is significantly mutated in lung adenocarcinoma. This project investigates the effects of the SMARCA4 mutation on alternative splicing and how it relates to tumorigenesis. We find that missense and non-missense mutations in SMARCA4 have different effects on splicing.

2.2 Introduction

While the human genome does not contain many genes compared to other species, it can achieve considerable complexity in part due to alternative splicing. Alternative splicing involves different permutations of the inclusion and exclusion of exons. Because of this, studying the regulation of alternative splicing can yield vast insights into any number of genetic diseases, such as cancer. Regulation of alternative splicing mainly involves factors that directly interact with exon-intron junctions. This regulation also includes the underlying sequence, such as the acceptor and donor splice sites, or proteins that interact with these sequences, such as splicing factors, such as U2AF1. However, the epigenetic state, such as nucleosome position, also affects alternative splicing. The proposed mechanism shows that as RNA Pol II transcribes pre-mRNA, parts of the epigenetic state, such as nucleosomes or DNA methylation, act as "speed bumps" and modify the elongation rate of RNA Pol II. This modulation of elongation has downstream consequences on the recruitment of splicing factors and can change which splicing events occur. Other work has also pointed toward SMARCA4 regulating the binding of splicing factors to RNA as it is being transcribed [16]. To investigate this further, we reanalyzed published RNA-seq data of lung adenocarcinoma patients with and without SMARCA4 mutations to understand its consequences on alternative splicing.

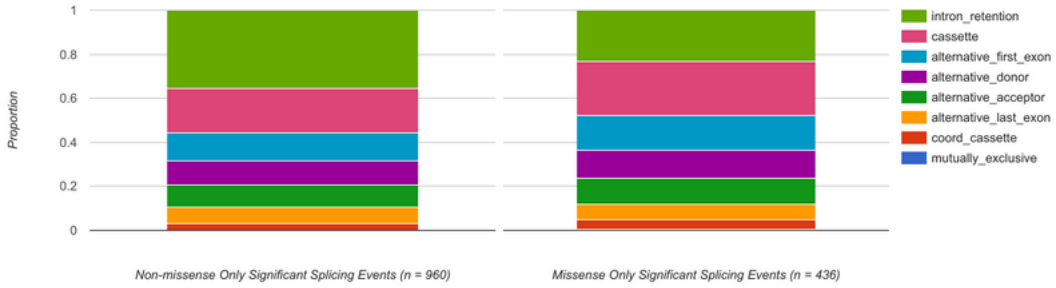

Figure 2.1: Significant splicing event counts between missense and non-missense mutations. The bar graph shows all types of splicing events for samples that contain either missense mutations in SMARCA4 or non-missense mutations in SMARCA4 that are significantly differentially spliced. Intron retention events are largest in non-missense mutations, whereas missense intron retention and cassette exon events are nearly equal.

2.3 Results

The output from the juncBase splicing quantification tool was used to quantify the different splicing events in the TCGA samples, shown in Figure 2.1 [7]. Of these, intron retention was found to be the largest splicing event in TCGA samples, with non-missense mutations at 36% of those in those samples. On the other hand, the number of cassette exon and intron retention events were nearly identical in the missense-only analysis at 107 and 102, respectively. Furthermore, when looking at the intron retention events between missense and non-missense groups, we find that only 36 events are shared.

2.4 Methods

Data was taken from 495 TCGA Lung Adenocarcinoma patients. Reads were aligned with HISAT2. The juncBase tool was used to quantify changes in splicing and

identify the types of splicing events. The wild-type for the comparisons contained no RBM10, U2AF1, or SMARCA4 mutations in the TCGA samples. Each sample was also annotated with whether they had low, high, or unknown expression of SMARCA4. Only the samples with a differentially-spliced corrected p-value less than 0.05 were retained. Only the known intron retention events were recorded when quantifying the splicing events in Figure 2.1.

Chapter 3

Add-seq: a novel method for inferring nucleosome positions on long-reads using angelicin-modified DNA

3.1 Abstract

Nucleosomes are DNA wrapped by histone octamers and regulate several biological processes. Nucleosomes regulate by preventing proteins and transcription factors from interacting with the bound DNA. Many sequencing methods locate the bound DNA by enriching sequences bound to nucleosomes. However, these methods can only infer one or two nucleosomes at once. Several nucleosomes coordinate together to influence protein binding. We developed Add-seq, which modifies nucleosome-free DNA with an angelicin small molecule. The small molecule creates DNA modifications on thymine bases and maps multiple nucleosomes on a single read in conjunction with nanopore

sequencing. We show that angelicin-modified DNA can be sequenced on the nanopore and, in the aggregate, can show patterns of accessibility consistent with chromatin biology. We identify key issues that angelicin has and suggest further developing this approach.

3.2 Introduction

Nucleosomes consist of a histone octamer that wraps approximately 147 bp of DNA and regulates transcription by occluding binding of transcription factors and polymerases [41, 30]. The typical histone octamer consists of H2A, H2B, H3, and H4 histone proteins, forming the protein complex that binds DNA [33]. Because of nucleosomes' ability for gene regulation, several methods have been developed to map the positions of nucleosomes. MNase-based approaches use the micrococcal nuclease enzymes, which cleave and chew away DNA not bound by the nucleosome [36, 9]. ATAC-seq uses Tn5 transposase to cleave DNA around accessible regions and enriches DNA bound by nucleosomes [8]. In these cases, the fragmented DNA is sequenced using a short-read sequencing platform like Illumina and reads mapped back to the reference genome. These reads give the position of single nucleosomes averaged across a large population of cells. Because the nucleosomes mapped by these methods only represent the average position of single nucleosomes. It can be challenging to infer whether a set of nucleosomes may have been positioned in sync or phased in a cell subpopulation. Sets of nucleosomes are positioned in discrete structures to have different gene regulation,

and current short-read methods cannot distinguish between the different structures [42].

In this paper, we present Add-seq, a method to observe these discrete structures of nucleosome positions. Our method combines two approaches to map nucleosomes on single molecules, 1) avoiding fragmentation by labeling DNA using a small molecule, and 2) reading the modified bases using nanopore sequencing.

The first part is to avoid fragmenting DNA by labeling DNA accessible through DNA modification. Our method uses a small molecule called angelicin. This furanocoumarin intercalates between DNA strands and will form a DNA modification on thymine nucleotides upon treatment by 365 nm UV light [27, 18]. This DNA modification will take place only on nucleotides not bound by nucleosomes. Thus, we can infer the position of nucleosomes if we know where the modifications are located. To validate our approach, we treat yeast with angelicin (3.1A). We perform seven rounds of modification, treating DNA in vivo and in vitro with angelicin and UV light multiple times to achieve good training data and higher resolution of nucleosome positions. We then purify the DNA and sequence it using ONT nanopore sequencing technology.

The second part is determining where the modifications are positioned using nanopore sequencing (Figure 3.1B). Nanopore sequencing involves a biological pore embedded in a lipid bilayer with an electrical current running across the pore [19]. As DNA is funneled through the pore with a motor protein, the kmers of the DNA cause shifts in the electrical current in a base-specific manner. This sequencing approach allows for sequencing long strands of DNA. This approach also allows for detecting DNA modifications, as the change in the chemical structure of the nucleotide due to the

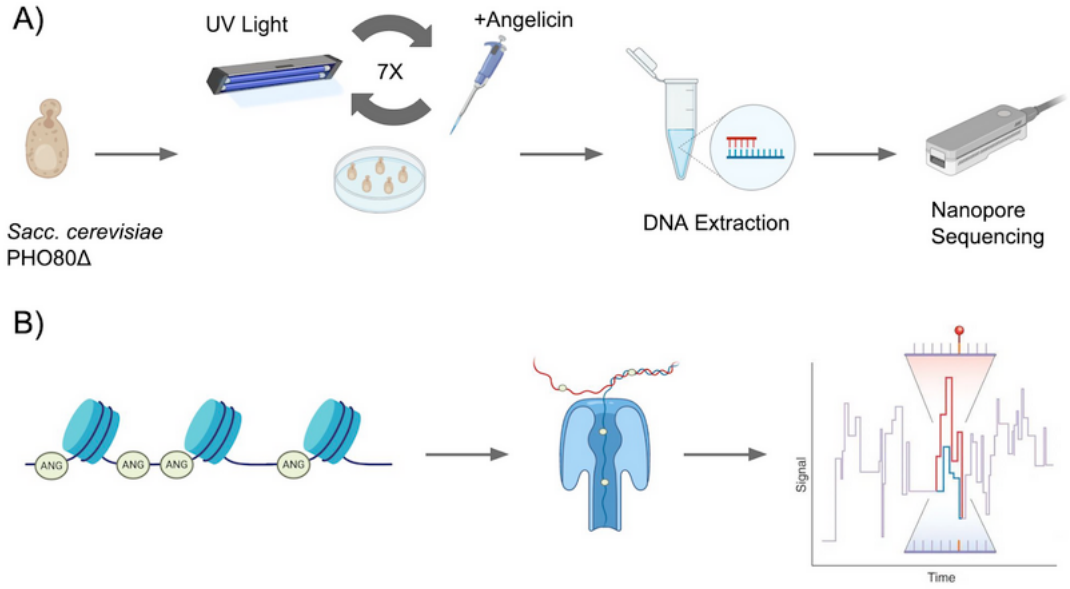

Figure 3.1: Add-seq uses angelicin and UV light to measure chromatin accessibility on long-reads with nanopore sequencing. A) Yeast cells are placed into a petri dish on ice. Angelicin is added to the mixture, and the cells are treated with UV light seven times. A high molecular weight DNA extraction purifies the treated DNA and sequenced by nanopore sequencing. B) Angelicin preferentially modifies accessible chromatin that isn't bound by nucleosomes. When the modified nucleotides are sequenced on the nanopore, changes in the electrical current can be used to determine where the modified base was positioned and, as a result, whether the chromatin was accessible around that position.

modification will be reflected in the electrical current [38, 46].

Our method will label accessible chromatin on long DNA strands and then detect the modifications while avoiding fragmentation using nanopore sequencing. To validate that our approach works, we sequence angelicin-modified DNA. We also use a meta-gene analysis across transcription start sites and orthogonal short-read-based methods to ensure we can observe known chromatin biology. Lastly, we find several difficulties with using an angelicin-based approach and offer several pathways towards

resolving those issues.

3.3 Results

3.3.1 Angelicin-treated DNA can be sequenced on the nanopore

We did a preliminary analysis using the PyMol structures to estimate whether it is likely that angelicin-modified DNA will be able to fit through the pore. Although we do not know the actual structure of the pore used in Oxford Nanopore flowcells, we looked at the pore structure we know was likely used to start. From this, we expect the DNA will fit through the pore.

To validate our approach, we performed nanopore sequencing on yeast DNA, one negative control treated with UV light, and one positive control treated with UV light and angelicin. Because UV light can cause DNA damage, such as thymine dimers and abasic sites, we treated our negative control with the same amount of UV light as the positive control. We aligned the sequencing data to the sacCer3 yeast genome reference using `minimap2` and aligned the signal data using `nanopolish eventalign`. Comparing the densities of the electrical current signal, we found that for kmers that contained the motif of interest, we didn't observe a shift in signal density; however, when we looked at kmers that did, we observed a secondary distribution. Previous papers have shown that modifications can cause shifts in the electrical current and used those to determine modification likelihood [38, 46]. Based on these results, we believed we could sequence DNA modified with angelicin with nanopore sequencing.

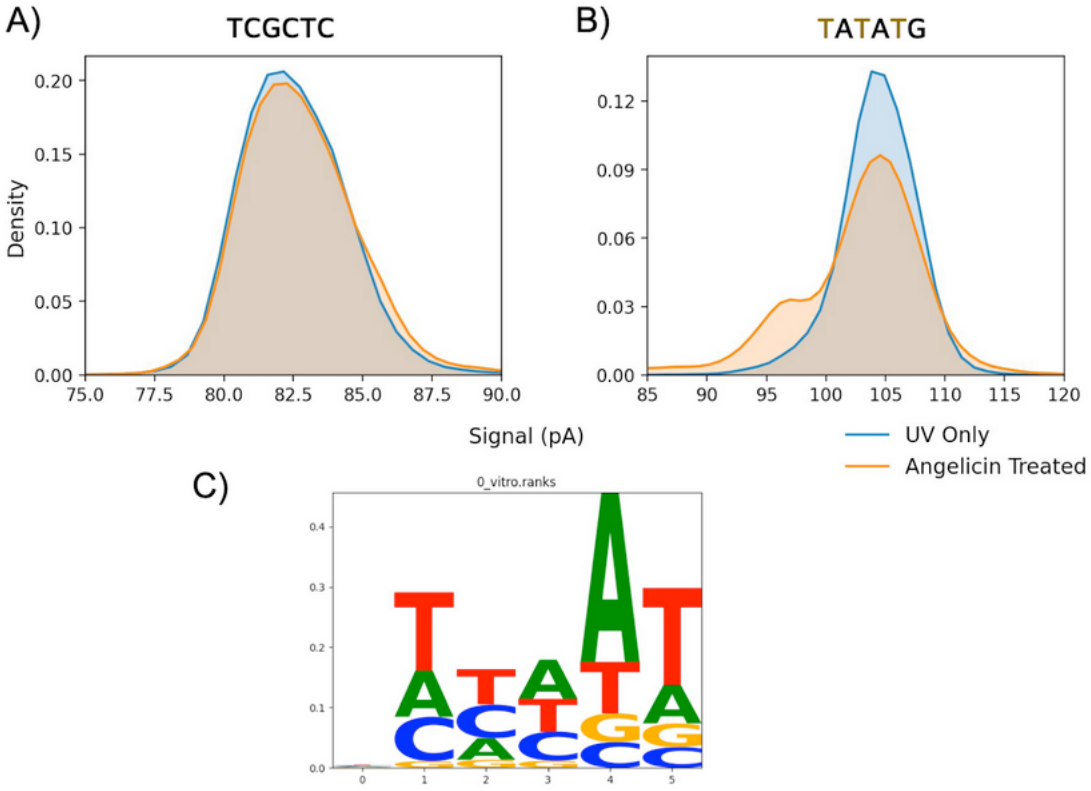

Figure 3.2: Signal distribution plots of modifiable kmers indicate sequencing of angelicin-treated DNA. Kernel density estimate plots of the nanopore signal data from yeast treated with UV light only (blue) and angelicin treated in vitro yeast DNA (orange) for two kmers, TATATG (left) and TCGCTC (right). B) The plot for the kmer TATATG, with modifiable nucleotides colored, shows a second distribution near 95 pA expected to come from modified nucleotides. A) The plot for the kmer TCGCTC, has no dinucleotides that are preferentially modified, and the kernel density estimate plots show that the signal distributions UV only and angelicin treated are very similar, indicating a lack of modification. C) Sequence logo of highest ranking kmers. Kmers were ranked based on the GMM for the positive control and single Gaussian using the KL divergence metric, and the 100 highest-scoring kmers were used to generate a sequence logo. The high level of TA and AT within the motif indicates that higher-ranking kmers tend to contain the motif associated with angelicin

We treated yeast DNA with 20 µM, 100 µM and 500 µM concentrations of angelicin and an untreated control sample. We made nanopore sequencing libraries from each treatment group and sequenced each over 72 hours. For the untreated, 20 µM, and 100 µM samples, we stopped sequencing after reaching 50,000 reads. We then sequenced the 500 µM library until sequencing finished. We observed that as the angelicin concentration increased, the throughput decreased significantly, taking longer to achieve 50,000 reads. Although sequencing

Following basecalling, we aligned signal current to the yeast genome using `nanopolish`. We then compared untreated and 500 µM angelicin concentrations by the signal distribution of TA and AT containing kmers. For these kmers, we observed the 500 µM kmer signal distribution often contained a bimodal distribution. The signal distribution would often be unimodal for kmers that did not contain TA or AT dinucleotides. The angelicin modification causes the secondary peak by influencing the signal measured current signal during nanopore sequencing.

3.3.2 Angelicin treated DNA shows patterns of nucleosome positioning

To evaluate our ability to detect angelicin-modified nucleotides, we followed a previously published paper approach using GpC methyltransferases [51]. We trained Gaussian mixture models on our control data, a single-component Gaussian for the negative control and a two-component Gaussian Mixture model for the positive control. These were trained for every 6-mer from the `nanopolish eventalign` signal alignments

average picoamp measurement. We used the Kulback-Leibler divergence by sampling to compare kmer models from the positive and negative controls. This metric gives a value where the higher the value, the more significant the difference between the two distributions being compared. We performed a sequence logo analysis on kmers with a KL divergence score greater than two (Figure 3.2C). The sequence logo preferred As and Ts in the kmers, corresponding to the positions we expect to modify by angelicin. Overall, this showed that across all kmers, we could find patterns that match the chemistry of how angelicin interacts with DNA.

We performed nanopore sequencing on nuclei treated with angelicin, so nucleosomes would not allow modifications, and we tested whether we could map nucleosomes in aggregate. We used the scoring approach from Wang et al. for the nuclei-treated sample. We scored each kmer in each read with the likelihood of being modified from the positive control Gaussian mixture model compared to the likelihood of being unmodified from the negative control Gaussian. To validate that the modification positions are biologically relevant, we looked at how likely modification occurs across several loci (Figure 3.3). The locus around a transcription start site shows a characteristic chromatin accessibility signal. Upstream of the transcription start site, the locus is generally accessible to allow for transcription factors and polymerases to bind to allow for transcription. Downstream, within the gene body, nucleosomes are packed next to each other, so overall accessibility is lower compared to linker regions, and a regular pattern of inaccessibility interspersed with accessible linker regions is expected.

Furthermore, the first nucleosome is expected to be the most well-positioned,

with subsequent nucleosomes less positioned going downstream. To see if angelicin-treated nuclei reflect this pattern, we averaged the modification score for a 1200bp window across every transcription start and end site in yeast (Figure 3.3A, B). The scores from this plot are flipped, so the more modification rate results in more negative scores. From this metagene plot, we found the scores followed the expected pattern, upstream having a wide area of negative scores. This correlates to our expectation of higher accessibility and lower modification rates within the gene body. Near the transcription termination site, there is a known pattern of accessibility surrounding this area [10]. Comparing our data here, we also saw a strong signal of accessibility closely around the transcription termination site.

In addition, we also compared scores with nucleosome maps in yeast [6]. This paper used modified histones to map nucleosomes across the yeast genome. When we averaged the scores across all nucleosomes in the paper, we found a pattern of inaccessibility and the center of the nucleosome, and then waves of accessibility and inaccessibility fanning out from the center of the nucleosome (Figure 3.3C). The Brogaard et al. paper also scored these nucleosome positions based on how consistently they were positioned [6]. Strong nucleosomes with high scores are often bound to this position, whereas weak nucleosomes with low scores are bound less often to the locus. When we separated nucleosomes based on a score cutoff to compare weak and strong signals, we found that while we maintained the signal we saw before with strong nucleosomes, the signal would disappear compared to weak nucleosomes. This further validates our approach because as a nucleosome is less often bound to a position, it is more likely that angelicin will be

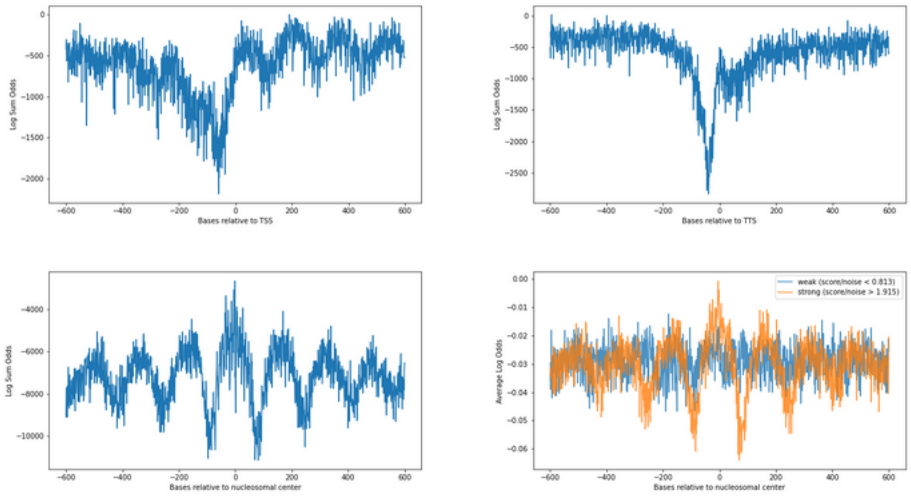

Figure 3.3: Genome-wide aggregate of modification probability reveals known chromatin accessibility patterns. For every transcription start site (TSS), transcription termination site (TTS), and nucleosome (citation here) in the yeast genome, the negative log-odds score for every base of every read that overlapped these regions were aligned. When comparing the sum of scores relative to the (A) TSS and (B) TTS, we see patterns typically associated with chromatin accessibility near the promoter region and lower accessibility in the gene body.

able to bind that locus. Overall, we saw that angelicin could modify linker DNA and read the signal closely matching results from orthogonal approaches.

3.3.3 Kmer-skipping causes difficulty in modification detection and nucleosome mapping

We observed that in positions where we expect modifications to occur, several large deletions span the position and several bases upstream and downstream of the position (Figure 3.4). These skipped kmers are positions where `nanopolish` could not assign signal current to sequence within the read. The bulky angelicin modification likely caused shifts in the electrical current that does not match any expected signal

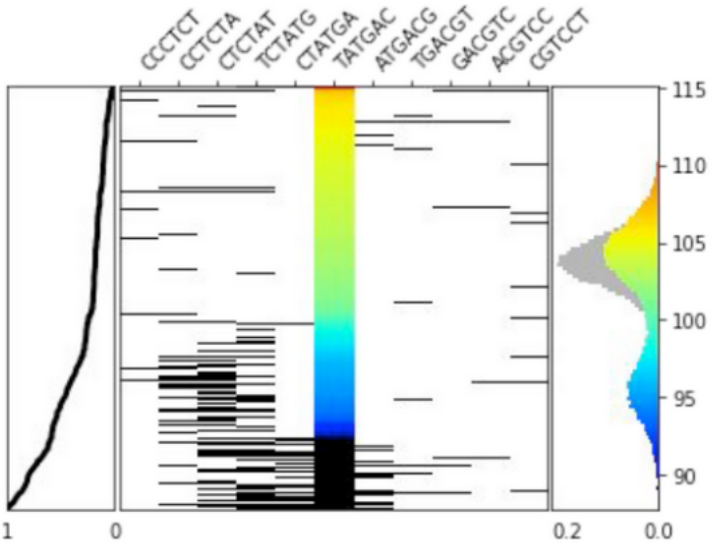

Figure 3.4: Angelicin modification has a high proportion of skipped kmers, yet observed data compares well with orthogonal data. Reads centered around the TATGAC position show many skipped kmers. The center column is colored by the current signal, as shown on the right of the image. The grey corresponds to the unmodified sample. Each row in the center represents a single read that maps to this locus. Black indicates the locus is skipped, meaning no signal is assigned to the position in `nanopolish eventalign` output. For reads with a skipped kmer at the TATGAC, where we expect modification to occur, surrounding kmers also tend to be skipped.

distribution of kmers, so the kmer is skipped. Because these kmers are skipped, our current pipeline cannot assign modification likelihood to the position. We have tried an approach that uses the frequency of kmer skipping to help with modification detection but did not find any significant difference in modification identification. From this, we have found that care needs to be taken when dealing with modifications that cause significant differences in the structure of the nucleotide, and further work needs to be done to identify these bulky modifications.

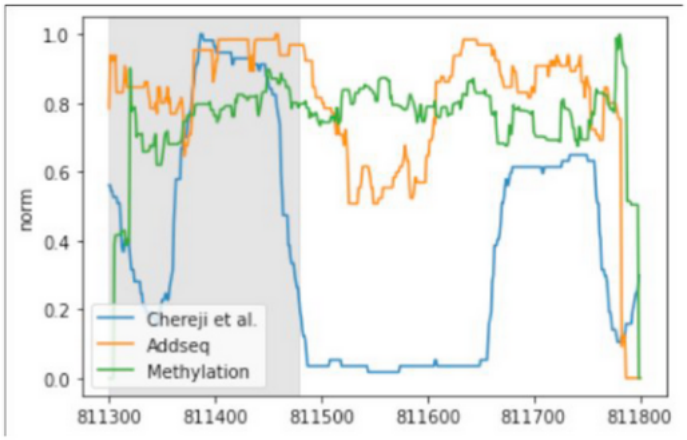

Figure 3.5: Orthogonal comparison indicates Add-seq is potentially better can be better than CpG methylation approaches. We compared the normalized score of chromatin accessibility from one sequencing track from Chereji et al. (blue) to the average nucleosome calls from Add-seq data (orange) and methylation-based data (green). In grey is highlighted the body of a gene in the yeast genome. We found that the Add-seq data correlated more closely to the orthogonal dataset than the methylation-based assay.

3.4 Discussion

While angelicin-based DNA modification can determine the positions of nucleosomes in aggregate, this approach has several challenges that need to be addressed. For machine learning-based approaches, having clean positive and negative controls are crucial for building models that accurately predict modifications from nanopore sequencing data. However, several aspects of angelicin prevent us from creating good positive controls compared to enzyme-based approaches. Because of the context of angelicin and bulkiness of the modification, we only expect to be able to modify one strand at a given time within a TA/AT context. As a result, we will not sequence the strand containing the modification half the time with standard nanopore sequencing.

Additionally, we tried to find third parties or companies able to synthesize DNA with angelicin modifications. To tackle this problem, we can increase the number of modified nucleotides sequenced by following an approach similar to 2D sequencing. This approach would ensure that the other strand is sequenced by joining the DNA strands via a hairpin adapter on one end. The hairpin-connected DNA would physically connect the two strands and allow the modification to be sequenced.

As we have attempted more sequencing runs with higher concentrations of modification, we have found that flowcell pores die more quickly, and as a result, more angelicin causes lower throughput. While preliminary analysis indicated that the pore should be able to accommodate the size of the modification, this result meant that angelicin-modified DNA might be interfering with the sequencing process. One hypothesis is that angelicin forms interstrand crosslinks, similar to its analog psoralen. Previous papers have shown that the chemistry of angelicin should not allow for the formation of crosslinks [3]. However, we decided to validate whether this was the case. Our gel results showed that a small fraction of reads forms interstrand crosslinks. These interstrand crosslinks will not be able to be broken during nanopore sequencing and thus force the pores to become clogged, reducing throughput. One way to alleviate this issue is by treating DNA with a base to break interstrand crosslinks [44]. Base treatment has been used with another furanocoumarin, psoralen, which generally forms DNA crosslinks.

Despite these challenges with using angelicin for resolving chromatin structure, there are still benefits to using it as a modification. Compared to enzyme-based approaches, angelicin is significantly cheaper for modifying the same amount of DNA.

Angelicin is an exogenous modification that does not naturally exist as a modification in cells. For GpC methyltransferases in genomes that endogenously have CpG modification, it cannot be easy to resolve whether the enzyme treatment made the mark. In comparison, we expect angelicin to interact differently compared to the other modifications during nanopore sequencing.

Furthermore, we compared our approach to GpC methyltransferase approaches with an orthogonal short-read approach (Figure 3.5). In Chereji et al., a chemical cleavage assay was used to get precise measurements of nucleosome positioning and compare the consistency with which a nucleosome would bind to a specific locus [11]. When comparing the modification scores across all of these nucleosomes, we found similar expected patterns of inaccessibility at the center of the nucleosome and regular intervals of accessibility and inaccessibility as DNA modified around nearby nucleosomes. In addition, we separated the nucleosomes based on scores provided by Chereji et al., corresponding to the consistency of nucleosome positioning at a given locus. For the well-positioned nucleosomes with high scores, we still observed the same pattern as before. However, for not well-positioned nucleosomes with low scores, we found that no regular intervals of accessibility could be found. With these results, we conclude that angelicin-modified DNA sequenced on the nanopore can successfully find patterns of chromatin accessibility that correlate to patterns found with short-read-based methods.

3.5 Methods

3.5.1 Angelicin treatment of DNA

DNA or cells were placed into Petri dishes with a layer of ice/cold water on the bottom to maintain cooler temperatures for the cells. Angelicin, with a concentration of $2\,\mathrm{mg\,mL^{-1}}$, was added to the cells. We allowed the angelicin to intercalate into the DNA for 5 minutes. Then the petri dish was treated with UV light at 365 nm UV light with a UV Stratalinker 2400 approximately 5 cm away from the light source. We repeated this process of angelicin treatment and UV light treatment for seven rounds. Ice was readded to maintain the temperature as needed. As a negative control, we also treated DNA with ethanol and subjected it to the same UV light treatment. For concentration experiments, DNA was treated with 0 µM, 20 µM, 100 µM, and 500 µM concentrations of angelicin and otherwise followed the same protocol as above.

3.5.2 Nanopore Sequencing

We took 1 µg of DNA for each and prepared DNA libraries with the SQK-LSK109 kit. We selected longer DNA fragments with Long Fragment Buffer. The sequencing runs were sequenced for 72 hrs. For concentration experiments, the 0 µM, 20 µM, and 100 µM libraries were run on the same flowcell until we achieved approximately 50,000 reads and the last run with the 500 µM was allowed to run until 72 hrs of sequencing had been performed.

3.5.3 Basecalling and preprocessing

We basecalled the data with `guppy` 2.3.7. For our concentration experiments, data was basecalled with `guppy` 6.1.7. We aligned reads to the sacCer3 genome with `minimap2` [31]. We then aligned signal data from the reads to the genome with `nanopolish` [46]. For our control data, we only used uniquely mapping reads for training.

3.5.4 Training of Gaussian mixture models

We followed a strategy similar to Wang et al. to train Gaussian mixture models [51]. For each 6-mer from our positive control, we trained a two-component Gaussian mixture model on the average signal measurement. For each 6-mer from the negative control, we fit a single Gaussian to the signal data distribution. We measured the difference between the signal distributions from the positive and negative control using the Kulback-Leibler divergence using sampling [17, 34].

3.5.5 Scoring algorithm

We scored each kmer by how likely it was to contain a modification relative to our negative control following the same approach from Wang et al. [51]. Briefly, for a given kmer and the corresponding signal measurement, the score is the likelihood of the signal measurement from the two-component Gaussian mixture model trained on the positive control, divided by the sum of the likelihood the signal measurement came from the GMM for the positive control and single Gaussian for the negative control.

Chapter 4

cawlr: a toolkit for analyzing chromatin accessibility With long-read sequencing data

4.1 Abstract

High-throughput mapping of nucleosome positions has primarily been done by enriching DNA fragments with short-read sequencing that maps only one or two nucleosomes simultaneously and is averaged across several cells. Several long-read chromatin accessibility methods have recently been developed to infer several nucleosomes on single molecules. However, the computational pipelines are method-specific, inflexible to new analyses, and challenging to run. We present `cawlr` (Chromatin Accessibility With Long Reads). This toolkit takes long-read sequencing data where accessible chromatin has been labeled via modification and infers nucleosome positions on single molecules.

`cawlr` implements a previous approach that uses dynamic alignment on modification probabilities to determine nucleosome positions. Our toolkit provides pipelines for training statistical models to call modifications and nucleosomes on nanopore data. It can also take inputs from other tools or technologies supporting modification BAMs.

Furthermore, our toolkit allows for visualizing kernel density estimates of control scores for estimating the accuracy of the toolkit, visualizing nucleosome calls on the UCSC Genome Browser, and clustering the calls to find distinct nucleosome structures. The code repository for this toolkit can be found at https://github.com/BrooksLabUCSC/cawlr-rs. This toolkit provides a platform for developing and evaluating new methods and tools to answer biological questions related to chromatin accessibility.

4.2 Introduction

Nucleosomes are protein-DNA complexes that provide an additional mechanism for gene regulation by exclusion of transcription factors and polymerases. Several methods have been developed for mapping the positions of nucleosomes genome-wide with high-throughput short-read sequencing [8, 23, 43]. These methods enrich DNA fragments that correspond to the position of a single nucleosome with high accuracy. These approaches only map single nucleosomes and don't allow for mapping them in the context of other nucleosomes and across the body of a gene. However, advances in long-read sequencing have enabled methods that can map several nucleosomes on a

single molecule. [51, 29, 47, 45, 1]. Like polymerases and transcription factors, nucleosomes can prevent DNA modification by preventing methyltransferases from binding [2]. These methods selectively modify nucleotides in the linker region, chromatin that is not bound by nucleosomes, using methyltransferases. These methods then infer the positions of nucleosomes based on where labeled nucleotides are located. However, each method requires a custom pipeline to analyze the data, and using these pipelines with custom analyses can be challenging.

In this paper, we present `cawlr`(chromatin accessibility with long reads), which provides a sequencing technology-agnostic toolkit for inferring nucleosome positions on single molecules (Figure 4.1A). Our method implements the computational approaches from Wang et al., which mapped nucleosome positions on single DNA molecules labeled by a GpC methyltransferase [51]. For the first step of modification detection from nanopore sequencing on r9.4.1 flowcells, we align nanopore signal measurements to the genome with `nanopolish` [46]. Then, we train Gaussian mixture models on the signal data for each kmer from the control data. While a 5-mer is usually measured in the pore, `nanopolish` will extend the kmer to allow for a more accurate association between signal and sequence. We then score every modifiable position for the likelihood of modification, compared to the control, using the Gaussian mixture model for that kmer. Lastly, we infer the most likely position of linkers and nucleosomes across the read using a hidden Markov model (HMM) and dynamic programming. Furthermore, we've expanded this approach to other modification detection tools and sequencing technologies by allowing BAM files that contain modification probabilities and positions using the ML and MM

tag, respectively. This allows us to expand the possible inputs to include outputs from newer Oxford Nanopore and Pacific Biosciences tools that support this type of BAM format.

Our tool can visualize the distribution of scores from positive and negative datasets, which estimates how reliable the single-molecule analysis will be (Figure 4.1B). The nucleosome and linker calls are output into a BED format which can then be visualized along with other datasets on the UCSC Genome Browser [24]. We also provide visualization for determining discrete nucleosome structures. We have generalized this approach so users can choose the modification motif; therefore, methods using other DNA modifiers can be analyzed with `cawlr`.

To illustrate the strength of our approach, we validate our method by looking at various loci using nanopore sequencing data from the Wang et al. study [51]. They used the M.CviPI methyltransferase, which labels accessible chromatin with 5-methylcytosine in a GpC context, and sequenced the reads using Oxford Nanopore sequencing. To demonstrate the flexibility of `cawlr`, we analyzed data from a study that labeled adenosines within accessible chromatin regions with the 6-methyladenine modification using the Hia5 methyltransferase and sequenced using the Pacific Biosciences platform [47]. Overall, `cawlr` provides a general approach to detecting nucleosomes on single molecules and will help accelerate research in epigenetics using long-reads.

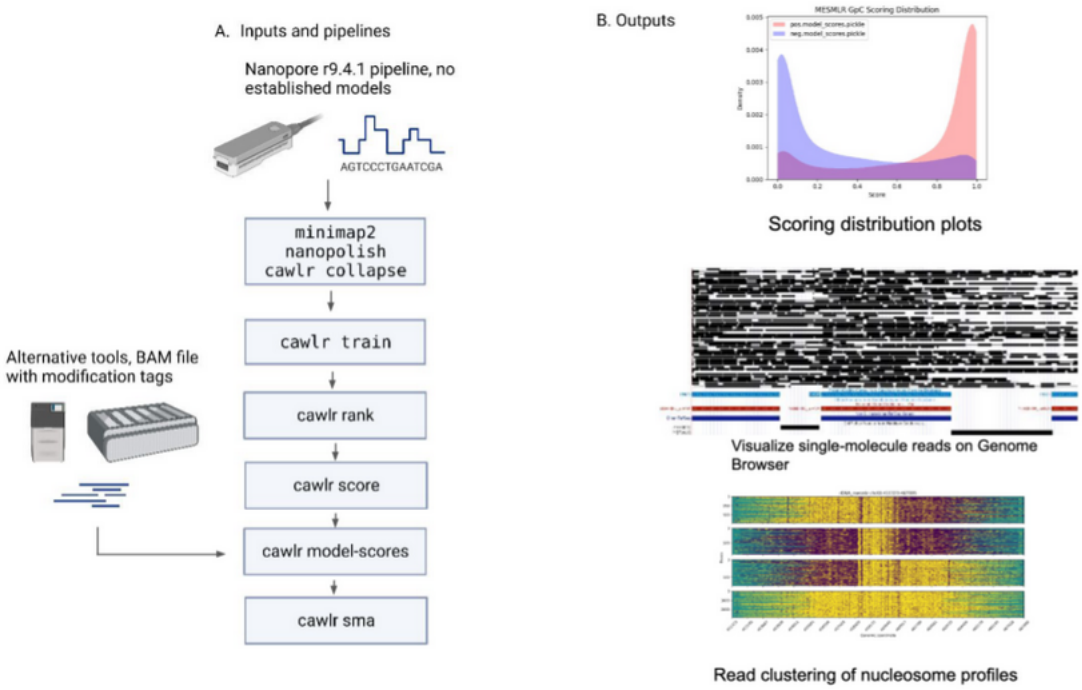

Figure 4.1: cawlr pipeline for analyzing modification data. A) The`cawlr` tool provides two pipelines. For training a model from scratch from nanopore sequencing on r9.4 flowcells, alignment is done with `minimap2` and `nanopolish`, and data is compressed using `cawlr collapse`. Gaussian mixture models are trained using `cawlr train` for each 6-mer. These Gaussian mixture models are ranked by KL divergence using `cawlr rank`. The likelihood of modification for each modifiable position is calculated with `cawlr score`. Kernel density estimates of the scores from the positive and negative controls are built using `cawlr model-scores`. And lastly, nucleosomes and linkers are called in each read using `cawlr sma`. For other datasets that contain modification information in BAM files with MM and ML tags, only the KDE of the controls are built, and single molecule analysis needs to be performed with `cawlr model-scores` and `cawlr sma`. B) `cawlr` provides various visualizations for evaluating chromatin accessibility. Our toolkit provides three outputs for visualizing the results. A scoring distribution density plots the scores from either workflow for comparison. Scores closer to one represent a higher likelihood of modification, and scores closer to zero represent a lower likelihood. A BED file is an output from `cawlr` sma that can be visualized on the genome browser. Each line represents a single read, with black boxes representing nucleosome positions and lines in between representing linker regions. Lastly, a script is provided that can take the BED file as input and outputs the reads cluster via K-means clustering. Each row represents a single read that maps to the locus input into the script. Yellow represents nucleosome positions, purple represents linker regions, and blue represents a position where the read didn't overlap. Each subplot represents a separate cluster, and the user can control the number of clusters.

4.3 Results

4.3.1 `cawlr` recapitulates nucleosome profiles from Oxford Nanopore sequencing data of GpC methyltransferase treated nuclei

To validate the tools, we used data from the MESMLR-seq data to look at single molecules aligned to the CLN2 locus in Saccharomyces cerevisiae (Figure 4.2A). The CLN2 gene encodes a cyclin that is involved in the regulation of the cell cycle in the G1 phase [12]. Because the CLN2 gene is transcribed during certain cell cycle phases, we expect heterogeneity in promoter region depending on the cell state. The previous Wang et al. study found three distinct clusters representing promoter accessibility as wide, narrow, and closed. We analyzed scoring data from Wang et al. NP-SMLR tool to ensure that our nucleosome calling works correctly. As shown in Figure 4.2A, we found the same three clusters as the previous study. This indicates we have successfully implemented the algorithm as described.

We focused on reads mapped to the rDNA locus to validate the tool further. In yeast, the rDNA locus is approximately 9kb long and consists of two regions that encode ribosomal protein genes and are repeated 100+ times on chromosome XII [50]. Each region inside the repeat can be transcriptionally silenced (activated/repressed) by making the chromatin inaccessible with high nucleosome density. Therefore, we expect reads that map to this locus to have four states for every combination of the active or inactive regions [13]. We processed the data using the `cawlr pipeline` from the raw FAST5 data. Controls were trained using `cawlr pipeline train-ctrls`. The in-vivo

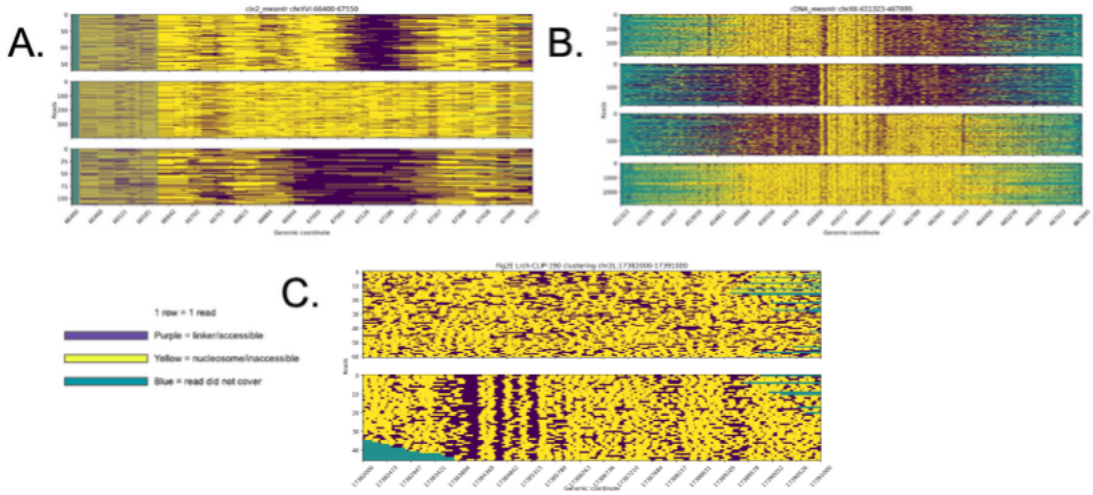

Figure 4.2: cawlr shows nucleosome profiles consistent with known biology. A) Analysis of MESMLR-seq data in CLN2 promoter region. Clustering with n=3 shows we can recapitulate the closed, narrow, and wide open promoter states. B) MESMLR-seq data aligning the rDNA locus shows four states, with each part of the rDNA subunit being randomly silenced via nucleosome positioning. C) Fiber-seq data from Stergachis et al. in the dual promoter Lrch-CLIP-190 locus through `cawlr sma`. We can recapitulate in the second cluster the accessibility in the promoter region.

treated data was then preprocessed using `cawlr pipeline preprocess-sample`, and rDNA locus analyzed using `cawlr pipeline analyze-region`. After running `cawlr` on reads that mapped to this locus and k-means clustering (k=4), we found four distinct clusters that correspond to each region being active or repressed (Figure 4.2C).

We also ran `cawlr` on sequencing data from the rDNA locus from another dataset from the MESMLR-seq paper, corresponding to a technical replicate. Interestingly, we did not find four distinct clusters compared to the previous replicate when we performed the same analysis. This result shows that variability between sequencing runs needs to be accounted for when analyzing chromatin accessibility with long-read sequencing technologies.

4.3.2 `cawlr` can infer nucleosome positions from different sequencing technologies and modification detection pipelines

To illustrate the tool's flexibility, use applied `cawlr` to data from the Fiber-seq paper, which uses Pac-Bio to detect modified bases and infer nucleosome position [47]. We used the `fibertools` pipeline to create BAM files containing modification scores [20]. We plotted the kernel density estimate of the modification likelihoods from the positive and negative controls (Figure 4.2D). The clear separation between the two distributions indicates the models for modification detection could accurately discriminate between modified and unmodified nucleotides. We then used our tool to analyze single molecules in the Lrch–CLIP-190 locus (Figure 4.2E). Our tool found nucleosome structures similar to those found by Fiber-seq. Overall, our method is flexible enough to enable analysis on various platforms.

4.4 Discussion

Calling nucleosomes on single molecules has been driven by being able to detect modifications with the latest sequencing platforms. While several protocols and modifications (or sets of modifications) have been used, there hasn't been a consensus on the best modification for nucleosome profiling. Furthermore, the statistical models used to detect modifications are constantly being developed. Because of this, we wanted to make sure the toolkit can be positioned to take inputs from a wide array of formats today and in the future. While we provide a pipeline that works with a particular

sequencing dataset, we can expand our approach to various techniques by taking data through BAMs with modification data.

This emphasis on using current data formats and not exclusively relying on our own should allow for combining nucleosome profiles with long-reads with other analyses. Other papers have already used this approach to examine endogenous CpG methylation with nucleosome profiles [1, 29]. We hope our toolkit expands the ability for other researchers to use those protocols as well. Long reads can also provide information about larger structural variations. This information is given "for free" alongside modification information in long-read sequencing. Although modifications can cause a decrease in sequencing accuracy, the combination of nucleosome profiling with structural variation can help put epigenetics in the context of more extensive genomic changes.

4.5 Methods

4.5.1 Data preprocessing

Nanopore sequencing data from Wang et al. was downloaded from accession PRJNA510813, and rebasecalled with `guppy` version 6.1.7 with the high-accuracy configuration [51]. Sequencing data was aligned to the sacCer3 reference genome using `minimap2` v2.24 [31]. For control data, only primary alignments were used. Signal events were aligned to the reference using `nanopolish` 0.13.3 [46]. These tools and their respective versions are also maintained in the provided Docker container.

4.5.2 Training GMM for modification detection

For the provided training models, `cawlr` adapts the pipeline used by Wang et al. [51]. Briefly, `nanopolish eventalign` is used to align signal data to currents. Signal data is filtered to be between 40 and 170 pA to avoid extreme current measurements that should not correspond to normal DNA kmers. The control datasets are then further filtered using DBSCAN for outliers. For the positive control dataset, a 2-component Gaussian Mixture Model for each kmer. For the negative control, only a single Gaussian is fit to the data. These models are trained on signal current measurements for every 6-mer. Ideally, there is 50,000X coverage across each kmer at minimum. To rank each kmer for the ability to distinguish modification probability, the Kulback-Leibler divergence score is calculated via sampling[17]. For a given position in the read that can be modified, the overlapping 6-mer with the highest divergence score is used to choose the set of positive and negative Gaussian Mixture Models to perform scoring. The GMM calculates the likelihood that a modified nucleotide produced a given signal measurement. Scores are calculated as the ratio of the likelihood of modification and the sum of the likelihood to be modified and unmodified. These likelihoods are only calculated for positions matching each modification's given motif. When passing the motifs to `cawlr model-scores` or `cawlr sma`, the motif string passed to the `--motif` parameter for GpC methyltransferase data such as from MESMLR-seq is `2:GC` for MESMLR-seq and `1:A` for 6mA data from Fiber-seq.

4.5.3 Nucleosome profiling and clustering

The inference of nucleosome positions follows the same approach as Wang et al. [51]. Briefly, for the controls, a kernel density estimate of the modification probability is created by sampling the score output. The density estimates are used to perform dynamic alignment of the scores to 147 base pair segments corresponding to the positions of nucleosomes or linkers that can be any number of bases long. The output for this module is a bed file compatible with the UCSC Genome Browser for visualizing the positions of nucleosomes on single reads. Clustering is performed via K-means clustering on reads that overlap specific regions. For reads that partially cover a locus, a dummy value of 0.5 is used. This plotting is performed using the `cluster_region.py` script in the `cawlr` git repo under `scripts/`.

4.5.4 Fiber-seq single-molecule analysis

The positive, negative control, and in vivo treated data from Stergachis et al. was downloaded for the Drosophila dataset [47]. Subread bams were converted to HiFi reads using `ccs` with the `--hifi-kinetics` parameter. Reads were aligned to the dm6 genome using `pbmm2`. The 6mA modifications were called and added to the bam files using `fibertools`. Kernel density estimates for the modification bams from the control data were generated with `cawlr model-scores`. The in-vivo data was then analyzed with `cawlr sma` with calls from the Lrch-CLIP-190 locus extracted and visualized with K-means clustering, n=2.

4.6 Discussion

Over the past few years, several papers have been published that use a variety of modification protocols and sequencing platforms. To date, there has not been a comprehensive comparison of each modification mark and the sequencing platform used for detection. We have outlined the issues with using an angelicin-based approach in chapter two. For GpC methyltransferase, the motif overlaps with genomes that contain endogenous CpG methylation. On the other hand, both ONT and PacBio have official and third-party tools and models for identifying the mark. For the Hia5 methyltransferase, it is not an endogenous nucleotide, however, it is not commercially available and lacks support from computational pipelines.

Furthermore, switching to r10.4.1 flowcells for ONT-based assays has made several pipelines obsolete. As a result, our toolkit has been developed to take several of these cases in mind. The regular pipelines can take nanopore sequencing data from r9.4.1 flowcells. It can also take BAM files containing modification information through the MM and ML tags. It has been generalized to allow for taking any set of motifs. As a result, our toolkit provides an analysis platform that should cover most methods used for measuring chromatin accessibility with long reads. One issue our pipeline does not address is managing DNA that contains several modifications at once. Our current pipeline focuses mainly on the yeast genome with no endogenous modifications. While other papers have explored the analysis of two modification marks at once, we expect that most methods will be unable to do more than that to achieve higher resolution or

map other DNA-binding proteins. As a result, using other methods that don't involve modification of nucleotides may be needed to get more information from single-molecule reads.

With these advancements, we expect long-reads to allow for understanding chromatin accessibility in various contexts. The DiMeLo-seq paper has used single-molecule long-read information to look at nucleosomes in repetitive regions in the human genome. Long-reads allow for spanning the repeat to some unique sequence and understanding how chromatin accessibility from CENP-A nucleosomes in those regions is regulated. Another layer of information long-reads provide is in understanding structural variation. Several tools have been developed to call larger structural variation events that short reads can not span. Furthermore, this structural variation can be phased into haplotypes, connecting these events across the chromosome.

Further work should be done to utilize this "free" information that long-read chromatin accessibility experiments provide to understand chromatin accessibility in the context of structural variation. Lastly, we hope to apply these techniques in cancer cell lines to address our original question of how recurrently mutated chromatin accessibility factors cause changes in alternative splicing. We have already shown in the first chapter that mutations in SMARCA4 have consequences in alternative splicing. And with these tools, we can see the nucleosome heterogeneity resulting from the mutations. The second part would combine this with long-read RNA sequencing and see the full isoforms produced by nucleosome heterogeneity. The ultimate goal would be to capture the epigenetics and splicing from an individual cell and tie these processes together directly.

This thesis represents an important step to make this goal achievable.

Bibliography

[1] Nicolas Altemose, Annie Maslan, Owen K Smith, Kousik Sundararajan, Rachel R Brown, Reet Mishra, Angela M Detweiler, Norma Neff, Karen H Miga, Aaron F Straight, and Aaron Streets. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. *Nature Methods*, 19(6):711–723, jun 2022.

[2] Andrew J Andrews and Karolin Luger. Nucleosome structure(s) and stability: variations on a theme. *Annual review of biophysics*, 40:99–117, 2011.

[3] M J Ashwood-Smith and E Grant. Conversion of psoralen DNA monoadducts in e. coli to interstrand DNA cross links by near UV light (320-360 nm): inability of angelicin to form cross links, in vivo. *Experientia*, 33(3):384–386, mar 1977.

[4] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, mar 2011.

[5] Oliver Bell, Vijay K Tiwari, Nicolas H Thomä, and Dirk Schübeler. Determinants and dynamics of genome accessibility. *Nature Reviews. Genetics*, 12(8):554–564, aug 2011.

[6] Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, jun 2012.

[7] Angela N Brooks, Li Yang, Michael O Duff, Kasper D Hansen, Jung W Park, Sandrine Dudoit, Steven E Brenner, and Brenton R Graveley. Conservation of an RNA regulatory map between drosophila and mammals. *Genome Research*, 21(2):193–202, feb 2011.

[8] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, dec 2013.

[9] Răzvan V Chereji, Terri D Bryson, and Steven Henikoff. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biology*, 20(1):198, sep 2019.

[10] Răzvan V Chereji, Josefina Ocampo, and David J Clark. MNase-sensitive complexes in yeast: Nucleosomes and non-histone barriers. *Molecular Cell*, 65(3):565–577.e3, feb 2017.

[11] Răzvan V Chereji, Srinivas Ramachandran, Terri D Bryson, and Steven Henikoff. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biology*, 19(1):19, feb 2018.

[12] F Cvrcková and K Nasmyth. Yeast g1 cyclins CLN1 and CLN2 and a GAP-like protein have a role in bud formation. *The EMBO Journal*, 12(13):5277–5286, dec 1993.

[13] R Dammann, R Lucchini, T Koller, and J M Sogo. Chromatin structures and transcription of rDNA in yeast saccharomyces cerevisiae. *Nucleic Acids Research*, 21(10):2331–2338, may 1993.

[14] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, may 2016.

[15] 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):461–465, jun 2010.

[16] Antoni Gañez-Zapater, Sebastian D Mackowiak, Yuan Guo, Marcel Tarbier, Antonio Jordán-Pla, Marc R Friedländer, Neus Visa, and Ann-Kristin Östlund Farrants. The SWI/SNF subunit BRG1 affects alternative splicing by changing RNA binding factor interactions with nascent RNA. *Molecular Genetics and Genomics*, 297(2):463–484, mar 2022.

[17] John R. Hershey and Peder A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on*

Acoustics, Speech and Signal Processing - ICA '07, pages IV–317–IV–320. IEEE, apr 2007.

[18] J E Hyde and J E Hearst. Binding of psoralen derivatives to DNA and chromatin: influence of the ionic environment on dark binding and photoreactivity. *Biochemistry*, 17(7):1251–1257, apr 1978.

[19] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239, nov 2016.

[20] Anupama Jha, Stephanie C Bohaczuk, Yizi Mao, Jane Ranchalis, Benjamin J Mallory, Alan T Min, Morgan O Hamm, Elliott Swanson, Connor Finkbeiner, Tony Li, Dale Whittington, William Stafford Noble, Andrew B Stergachis, and Mitchell R Vollger. Fibertools: fast and accurate DNA-m6A calling using single-molecule long-read sequencing. *BioRxiv*, apr 2023.

[21] Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews. Genetics*, 10(3):161–172, mar 2009.

[22] Steven M Johnson, Frederick J Tan, Heather L McCullough, Daniel P Riordan, and Andrew Z Fire. Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. *Genome Research*, 16(12):1505–1516, dec 2006.

[23] Theresa K Kelly, Yaping Liu, Fides D Lay, Gangning Liang, Benjamin P Berman, and Peter A Jones. Genome-wide mapping of nucleosome positioning and DNA

methylation within individual DNA molecules. *Genome Research*, 22(12):2497–2506, dec 2012.

[24] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, jun 2002.

[25] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews. Genetics*, 11(5):345–355, may 2010.

[26] Mirang Kim and Joseph Costello. DNA methylation: an epigenetic mark of cellular memory. *Experimental & Molecular Medicine*, 49(4):e322, apr 2017.

[27] Jun-ichiro Komura, Hironobu Ikehata, Yoshio Hosoi, Arthur D. Riggs, and Tetsuya Ono. Mapping psoralen cross-links at the nucleotide level in mammalian cells: suppression of cross-linking at transcription factor- or nucleosome-binding sites†. *Biochemistry*, 40(13):4096–4105, apr 2001.

[28] RD Kornberg and JO Thomas. Chromatin structure; oligomers of the histones. *Science*, 184(4139):865–868, may 1974.

[29] Isac Lee, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J Sedlazeck, Kasper D Hansen, Jared T Simpson, and Winston Timp. Simultaneous profiling of chromatin accessibility and methylation on

human cell lines with nanopore sequencing. *Nature Methods*, 17(12):1191–1199, dec 2020.

[30] Bing Li, Michael Carey, and Jerry L Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, feb 2007.

[31] Heng Li. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, oct 2021.

[32] P T Lowary and J Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology*, 276(1):19–42, feb 1998.

[33] K Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, sep 1997.

[34] Carolin A Müller, Michael A Boemo, Paolo Spingardi, Benedikt M Kessler, Skirmantas Kriaucionis, Jared T Simpson, and Conrad A Nieduszynski. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nature Methods*, 16(5):429–436, may 2019.

[35] Timothy W Nilsen and Brenton R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, jan 2010.

[36] M Noll. Subunit structure of chromatin. *Nature*, 251(5472):249–251, sep 1974.

[37] Alexander Payne, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. BulkVis: a graphical viewer for oxford nanopore bulk FAST5 files. *Bioinformatics*, 35(13):2193–2198, jul 2019.

[38] 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, apr 2017.

[39] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, dec 2011.

[40] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute*, 13(5):278–289, oct 2015.

[41] Timothy J Richmond and Curt A Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, may 2003.

[42] Ronen Sadeh and C David Allis. Genome-wide "re"-modeling of nucleosome positions. *Cell*, 147(2):263–266, oct 2011.

[43] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, mar 2008.

[44] Y B Shi, H P Spielmann, and J E Hearst. Base-catalyzed reversal of a psoralen-DNA cross-link. *Biochemistry*, 27(14):5174–5178, jul 1988.

[45] Zohar Shipony, Georgi K Marinov, Matthew P Swaffer, Nicholas A Sinnott-Armstrong, Jan M Skotheim, Anshul Kundaje, and William J Greenleaf. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nature Methods*, 17(3):319–327, mar 2020.

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

[47] Andrew B Stergachis, Brian M Debo, Eric Haugen, L Stirling Churchman, and John A Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498):1449–1454, jun 2020.

[48] Alison D Tang, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals down-regulation of retained introns. *Nature Communications*, 11(1):1438, mar 2020.

[49] B van Steensel, J Delrow, and S Henikoff. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature Genetics*, 27(3):304–308, mar 2001.

[50] J Venema and D Tollervey. Ribosome synthesis in saccharomyces cerevisiae. *Annual Review of Genetics*, 33:261–311, 1999.

[51] Yunhao Wang, Anqi Wang, Zujun Liu, Andrew L Thurman, Linda S Powers, Meng Zou, Yue Zhao, Adam Hefel, Yunyi Li, Joseph Zabner, and Kin Fai Au. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Research*, 29(8):1329–1342, aug 2019.