9-16-2019

Chromatin digestion by the chemotherapeutic agent Bleomycin produces nucleosome and Transcription Factor footprinting patterns similar to Micrococcal Nuclease

Joshua Michael Stolz
Rowan University

Recommended Citation

Stolz, Joshua Michael, "Chromatin digestion by the chemotherapeutic agent Bleomycin produces nucleosome and Transcription Factor footprinting patterns similar to Micrococcal Nuclease" (2019). *Theses and Dissertations*. 2736.
https://rdw.rowan.edu/etd/2736

CHROMATIN DIGESTION BY THE CHEMOTHERAPEUTIC AGENT BLEOMYCIN PRODUCES NUCLEOSOME AND TRANSCRIPTION FACTOR FOOTPRINTING PATTERNS SIMILAR TO MICROCOCCAL NUCLEASE

by
Joshua Stolz

A Thesis

Submitted to the
Department of Molecular and Cellular Biosciences
College of Science and Mathematics
In partial fulfillment of the requirement
For the degree of
Master of Science in Bioinformatics
At
Rowan University
May 18, 2019

Thesis Chair: Benjamin R Carone, PhD

Acknowledgements

I would like to thank Rowan University for its dedication to student research. Specifically, I would like to thank Doctor Benjamin Carone for taking the time to supervise this project and be the chair of my thesis. I would also like to thank Dr. Mary Alpaugh and Dr. Stephanie Spielman for taking the time to be on my committee. I would like to thank the Carone lab for processing and producing the libraries, specifically Elizabeth Richards. I would like to thank Ronak Dave and Adam Jamnik for doing the preliminary work necessary to execute this thesis.

Abstract

Josh Stolz
CHROMATIN DIGESTION by THE CHEMOTHEROPEUTIC AGENT BLEOMYCIN PRODUCES NUCLEOSOME AND TRANSCRIPTION FACTOR FOOTPRINTING PATTERNS SIMILAR TO MICROCOCCAL NUCLEASE
2018-2019
Ben Carone, Ph. D
Master of Science in Bioinformatics

Bleomycin (BLM), a glycopeptide antibiotic commonly used in chemotherapeutic treatments, has been shown to produce single and double stranded DNA breaks. Subsequent analysis of DNA fragmentation patterns has demonstrated preferential digestion of chromatin in the TSS of active genes and the ability to produce nucleosome-sized fragments within intact chromatin. Nucleosome positioning plays a critical role in the regulation of gene activation. Currently, micrococcal nuclease (MNase) is used as the standard for mapping the position of nucleosomes in the genome. In order to identify whether BLM can be used as an effective nucleosome-mapping agent, BLM was used to digest chromatin in *S. cerevisiae*, followed by Next Generation Sequencing of paired-end DNA fragments. Our results demonstrate comparable DNA fragmentation patterns for both nucleosomes as well as other DNA-protein interactions and furthermore explain the propensity for BLM to digest within the promoters of active genes. Finally, we show that BLM can be used to identify genome-wide nucleosome and Transcritption factor footprints as an effective alternative for MNase and additionally lacks the strong sequence biases of MNase digestion.

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

List of Figures

Figure Page

List of Figures (Continued)

Figures Page

Chapter 1

Introduction

The Structure of DNA

Deoxyribonucleic Acid (DNA) is a molecule that stores the information necessary for the inheritance of traits. In the famous Hershey-Chase experiment it was discovered that DNA, not Protein, is the molecule responsible for inheritance [1]. Ultimately, the ability of DNA to pass on information from one generation to the next is derived from the nature of its structure. Principally, DNA exists as a double helix. This structure was originally identified by Watson and Crick using X-ray crystallography and has been furthermore confirmed atomic force microscopy **(Figure 1)**. This double helix consists of a major groove, which can be seen with red arrows and a minor groove shown in blue arrows. [2] This structure arises from the anti-parallel structure of DNA strands in a double helix and serves to provide a specific surface available for DNA binding proteins. At the chemical level, DNA is comprised of a phosphate backbone a sugar and a nucleotide base. Watson and Crick also found that it was the composition of nucleotide bases that pair as Adenine and Thymine, as well as Guanine and Cytosine. [3] The ordering and composition of these nucleotides in the DNA strand can be interpreted by molecular machinery to carry out the central dogma including the transcription of DNA into RNA and the translation of RNA into Protein, the molecules responsible for cellular replication and function.

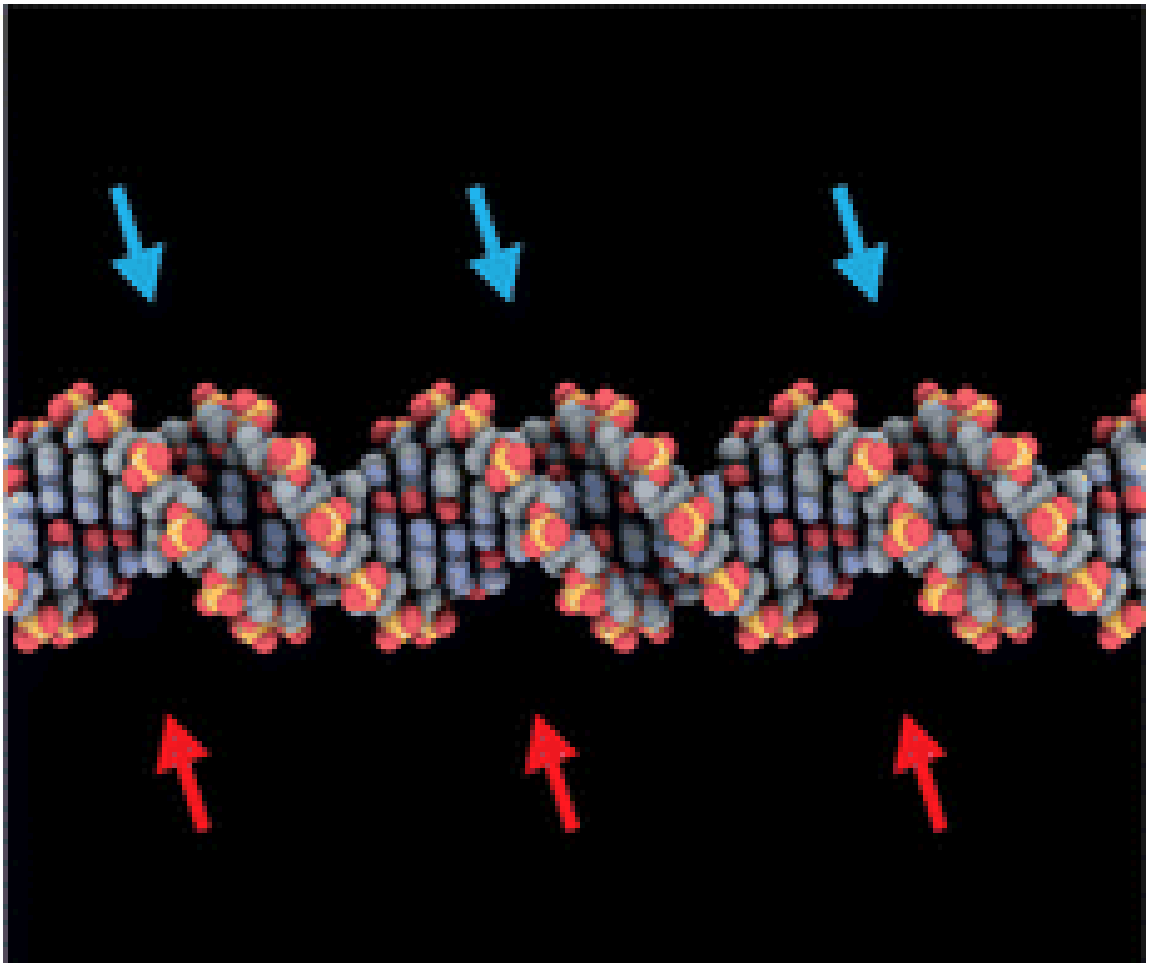


Figure 1. DNA Double Helix. DNA model showing the major groove of DNA with red arrows and the minor groove with blue. [2]

RNA Structure and Function

In the process of transcription, DNA is converted to a less stable nucleic acid counterpart, RNA. The first step in transcription is the binding of transcription factors to a promoter region, and it is at this point that gene is considered “poised”. This transcription factor then recruits RNA polymerase to bind to the transcriptional start site (TSS). RNA polymerase can create what is called a transcription bubble, which is the separation of the two DNA strands in the region of transcription. RNA is then formed by the addition of complimentary ribonucleic acids to the DNA template strand with the exception of thymine being replaced with uracil. The sugar phosphate backbone is then formed between the adjacent the complementary bases to create an intact RNA strand. Following the addition of all the complementary nucleotides, the completed RNA strand is released and may be further modified by a process called splicing. Splicing is the removal of introns that gives the cell the ability to create multiple versions of protein from a single gene. Once the transcript is modified it may move out into the cytosol through a nuclear pore. The structure of RNA is DNA except for in three cases: the deoxyribose on DNA is a ribose in RNA, RNA is a single stranded molecule, and Uracil is used instead of Thymine. This difference in structure is significant in that it reduces the half-life of RNA. [4] The reduced half-life of RNA has two major metabolic functions. First it allows the cessation of metabolic response to environmental conditions once homeostasis is returned. The second is it serves as a fail-safe in replication error so that when RNA is altered it ultimately degrades avoiding biological consequence. RNA serves as a temporary and decay mediated messenger for the cell.

Protein Structure and Function

Upon entering the cytosol from the nucleus, the process of creating an amino acid chain from RNA is started. A three base pair window in an open reading frame corresponds to a specific amino acid. This process, known as translation, is initiated when a tRNA binds to the start codon on mRNA. Each subsequent tRNA elongates the peptide chain by adding an amino that is complimentary to the codon. Amino acid side-chains contain unique properties and will contribute to both function and folding of the amino acid chain. Protein folding is the process by molecular physics of amino acid chains obtaining their 3-D structure. It is the intermolecular of amino-acid side chains that drive a polypeptide chain to secondary structure. [5] Alpha-helixes and beta-sheets are typical of secondary structure as they are stabilized by intramolecular hydrogen bonding. It is important to remember that this process is happening in an aqueous environment, and that is why tertiary structures are driven by hydrophobic and hydrophilic interactions and will often form a globular structure [6]. These proteins can form a larger protein complex called a quaternary structure comprised of more than one amino acid chain subunit. An example of a quaternary structure is hemoglobin with four protein subunits [7]. The degree to which genes are expressed regulates RNA levels which in turn regulate protein levels. This creates a system for carrying out and controlling cellular functions.

Epigenetic Mechanisms

Somatic cells descend from the same genome, however based upon cell signaling and gene regulation cells have the capacity to create pattern of expression that give rise to specialized functions within the body. This process is exemplified in the development of mammalian embryos. Early in embryo development cells are pluripotent and genes are largely expressed for transcription factors while genes with specific function within the whole organism experience gene silencing. As development proceeds cells obtain unique patterns of expression and lose the ability to be pluripotent. [9] For the persistence of this phenomenon to develop in living systems, cells require the ability to develop a “memory” through a specialized molecular mechanism called “Epigenetics”. This term was originally coined by Conrad Waddington who would describe a “ghost-like” phenomenon that hovered above genes and allowed them to remember their function and purpose [10]. This cellular memory is implemented by a variety of factors including the positioning of both nucleosomes and transcription factors within the genome. [11]

Nucleosomes

A nucleosome is a combination of DNA and eight histone proteins that contribute to the organization of DNA within chromatin. [12] The nucleosome is 146 base pairs of DNA wrapped around a histone octamer. [13] There are four different subunits of histones; H2A, H2B, H3, H4 **(Figure 2)**. [14] Nucleosomes functions to provide an organized compact structure for chromatin as there are six feet of DNA in each cell that has be organized and compressed to both prevent tangling and preserve space. As nucleosomes become more densely positioned gene expression decreases. During cell division DNA will move from "beads on a string" to one tightly packed chromosome. [15] When chromatin is conformed into densely compact nucleosomes, this is referred to as heterochromatin and typically has decreased gene expression **(Figure 3)**.

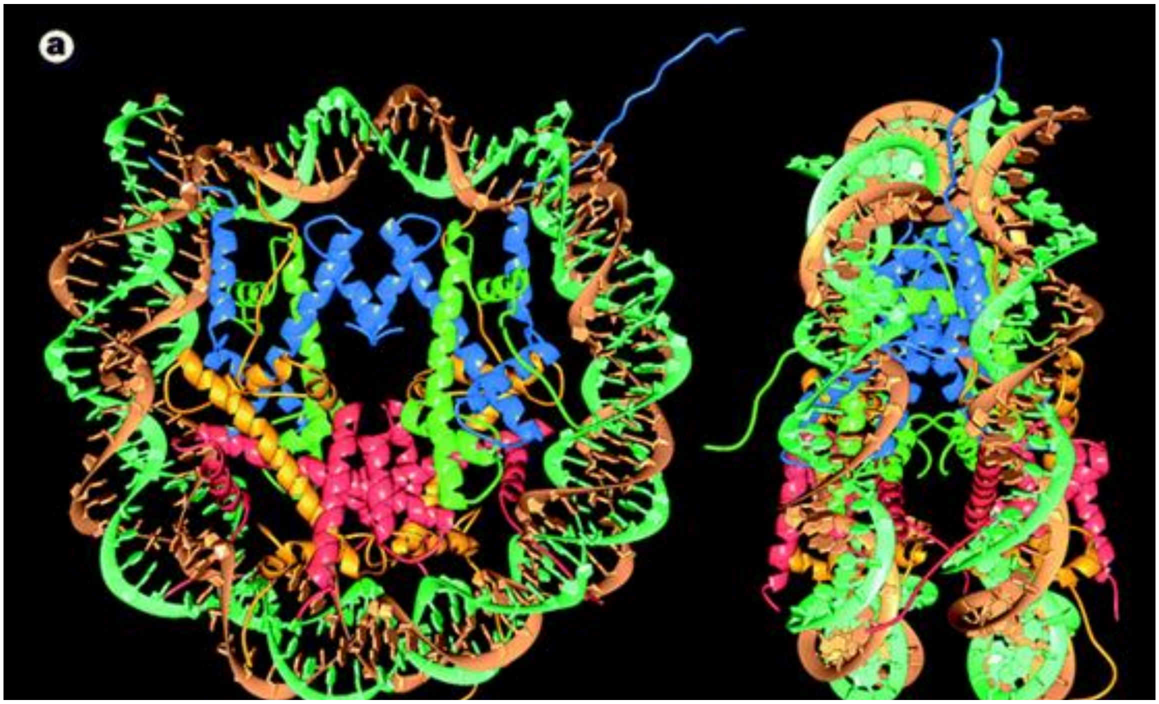


Figure 2. Nucleosome core particle. Ribbon representation of the eight histones (blue: H3; green: H4; yellow: H2A; red: H2B.) and 146 base pairs (brown and turquoise) associated with a nucleosome. [13]

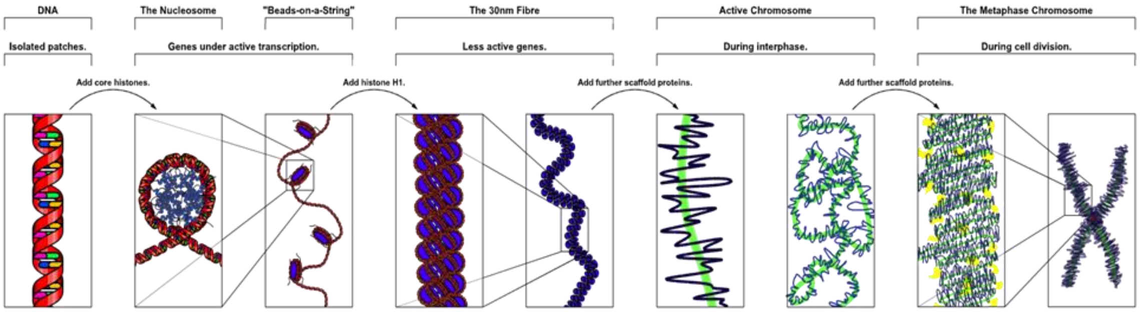


Figure 3. The current chromatin compaction model is displayed. Seen on the left naked DNA, followed by the addition of nucleosomes. Theses "beads on a string" are compacted as shown above to make a dense chromosome. [15]

The Role of Nucleosomes in Epigenetic Patterning

The role of the nucleosome in epigenetic processes is fundamentally established by both its location on DNA as well as its associated post-translational modifications. Histones are studded with sites for possible modification on its surface. [16] Within this modification system there are at least eight different classes of modification that occur in response to environmental changes. Being that the modifications to histone tails can have either hydrophobic or hydrophilic properties, nucleosomes will reorient themselves to reduce entropy in the system by either spreading farther apart or closer together. [17] In the case of chromatin condensing into a metaphase chromosome, histones are primarily modified by phosphorylation. As in many biological systems, nucleosome and histone modification are often dynamic, meaning the process is enzymatic and reversible. In the field of epigenetics, it is well understood that there is an immediate metabolic implication to nucleosomes positioning in that when nucleosomes are present in a promoter region the gene is significantly suppressed. [18] If a nucleosome is methylated and it is in a promoter region it will suppress expression. Methylation in this context is the attachment of a methyl group to a histone tail. If it is present in the coding region and methylated it will increase expression. This is countered by adding an acetyl group to the histone tail tends to make nucleosomes further apart leaving more room for DNA binding proteins. Nucleosome positioning within the promoter region is negatively correlated with gene expression **(Figure 4)**. The patterns of nucleosome occupancy and RNA Polymerase II occupancy can be seen in four separate clusters of the data (**Figure 4 A**). These data were clustered by gene expression levels and represent patterns of chromatin as it relates to regulating gene expression at these respective levels. It shows that after an active transcriptional start site is often a well-positioned nucleosome. These plots also show that in highly expressed gene, RNA Polymerase II is often positioned within the open reading frame. In each gene as the nucleosome depleted region (NDR) is elongated the gene expression increases **(Figure 4B).** [19] We can also see downstream of the transcriptional start sight that the nucleosome positioning decays in a stochastic model, meaning nucleosome are less well positioned the further downstream from a transcriptional start sight (TSS). [20]

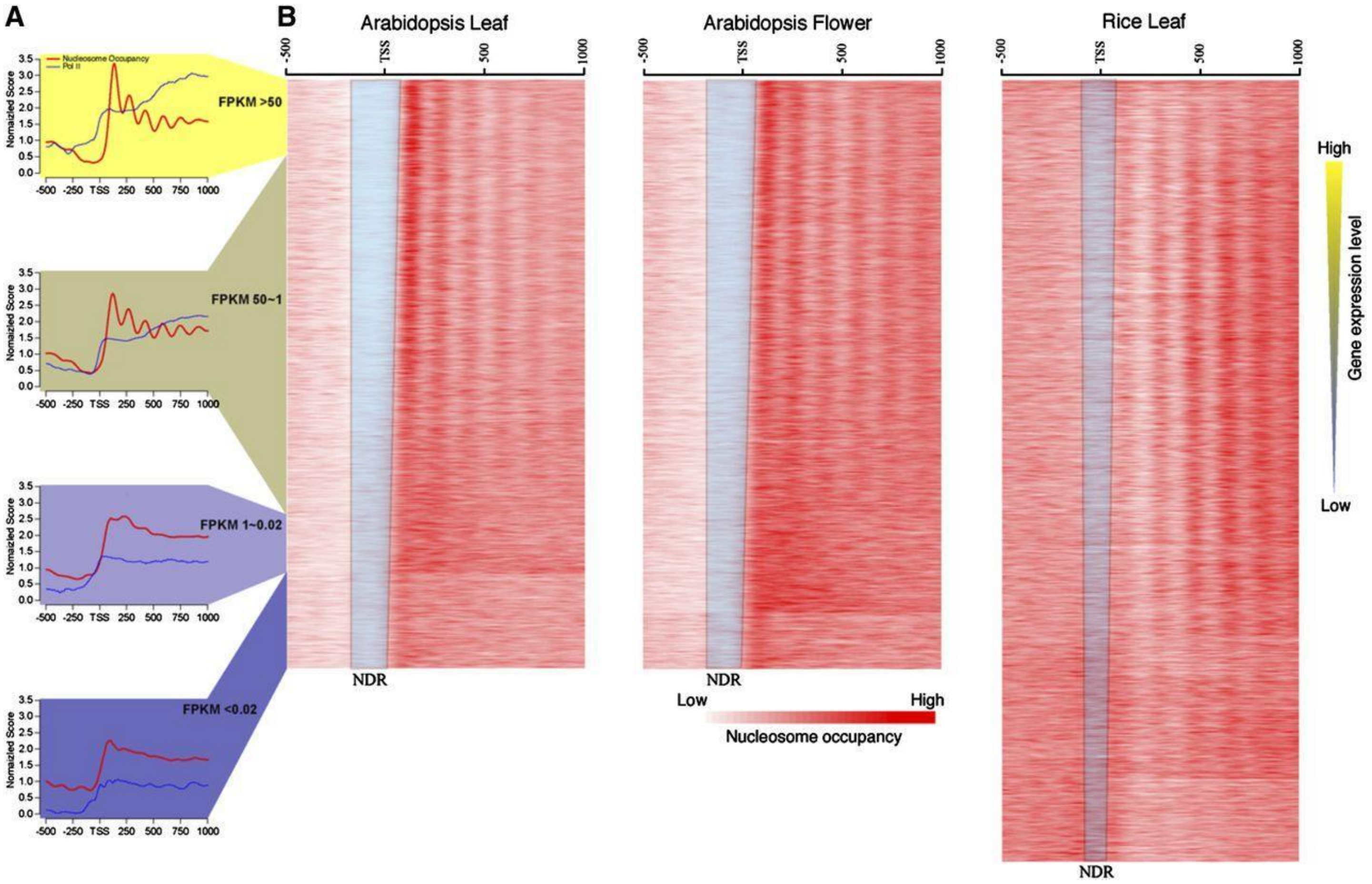


Figure 4. Relationship between gene expression and nucleosome positioning. A. Profiles of DNA Polymerase II (Pol II) and nucleosome positioning in subsets of the Arabidopsis leaf data set in blue and nucleosome occupancy in red. B, Heat maps of nucleosome position in Rice and Arabidopsis tissues. The blue trapezoid labeled nucleosome depleted region (NDR) on each figure is the nucleosome depleted region. RNA-seq data was obtained from the same rice tissues used for MNase-seq. [24]

The Role of Transcription Factors in Gene Expression

Transcription factors are common DNA elements associated with gene expression. Transcription factors act in the initiation of transcription by increasing the affinity to a promoter. A transcription factor is a protein that will bind to the promoter region of a gene and have an impact on the activation or suppression of that gene. It is well established that binding to DNA is a required property of transcription factors, however this is not enough to alter the activity of the gene. The transcription factor will likely also have an affinity for RNA polymerase as well. A single transcription factor can bind to multiple genes causing the activation of an entire pathway. For example, the GAL1 transcription factor activates all galactase genes giving a cell a simple mechanism to respond to environment. [21] The modification and production of transcription factors occurs in response to cell signaling and it is known that many transcription factors cannot bind to DNA without co-factors, and as a resulted are said to be modulated. [22] It has been recently discovered that some transcription factors will regulate DNA three-dimensional structure into loops in a similar way as CTCF (also known as 11-ziznc finger protein) binding proteins will. The YY1 gene has shown the ability to bind a both promoter and enhancer regions in the genome giving it the ability to reshape DNA structure. It can bind with RNA but also has an affinity for itself giving it the ability to bind to itself while bound to the genome and stabilizing its structure. [23]

CPG Islands Role in Epigenetics

A CpG site is a position in the genome where cytosine is followed by guanine from 5’ to 3’. This site is of importance because it is a common place for DNA methylation, which is a mechanism of gene suppression. 70-80% of the cytosine in CpG sites in mammalian species are methylated. [24] Regions that are over 50% CpG sites are referred to as CpG islands and are instrumental in regulation of gene expression as 70% of promoter regions are located near CpG islands. [25] A study was done on the hyper-methylation of CpG sites in colorectal cancer and it found that 1734 CpG islands were hyper-methylated in the cancer tissues when compared to healthy tissues. Methylation patterns in leukemia patients have given the ability to doctors of predicting medication response and early prognosis. The activation of genes gives promise that when cancer antigens are expressed medications can be used in collaboration with epigenetics [26]

S. cerevisiae as a Model to Study Epigenetics

When developing novel methods in epigenetics, the scientific field typically prefers *S. cerevisiae,* budding yeast as a model organism due to the low cost and short generation time. Extensive research on yeast has served to map the DNA elements and apply these concepts to analogous organisms. Whole genome sequencing of nucleosome acetylation in yeast has been performed to confirm nucleosome positioning and patterns. This data was then used to confirm the previously stated correlation between gene expression and the presence of nucleosomes in the promoter region. Transcription factor positioning is negatively correlated with nucleosome positioning **(Figure 5)**. This can be attributed to the need for transcription factors to bind to the promoter. [27] The foundational level of academic research has already been done on yeast and the low cost makes this an ideal model organism.

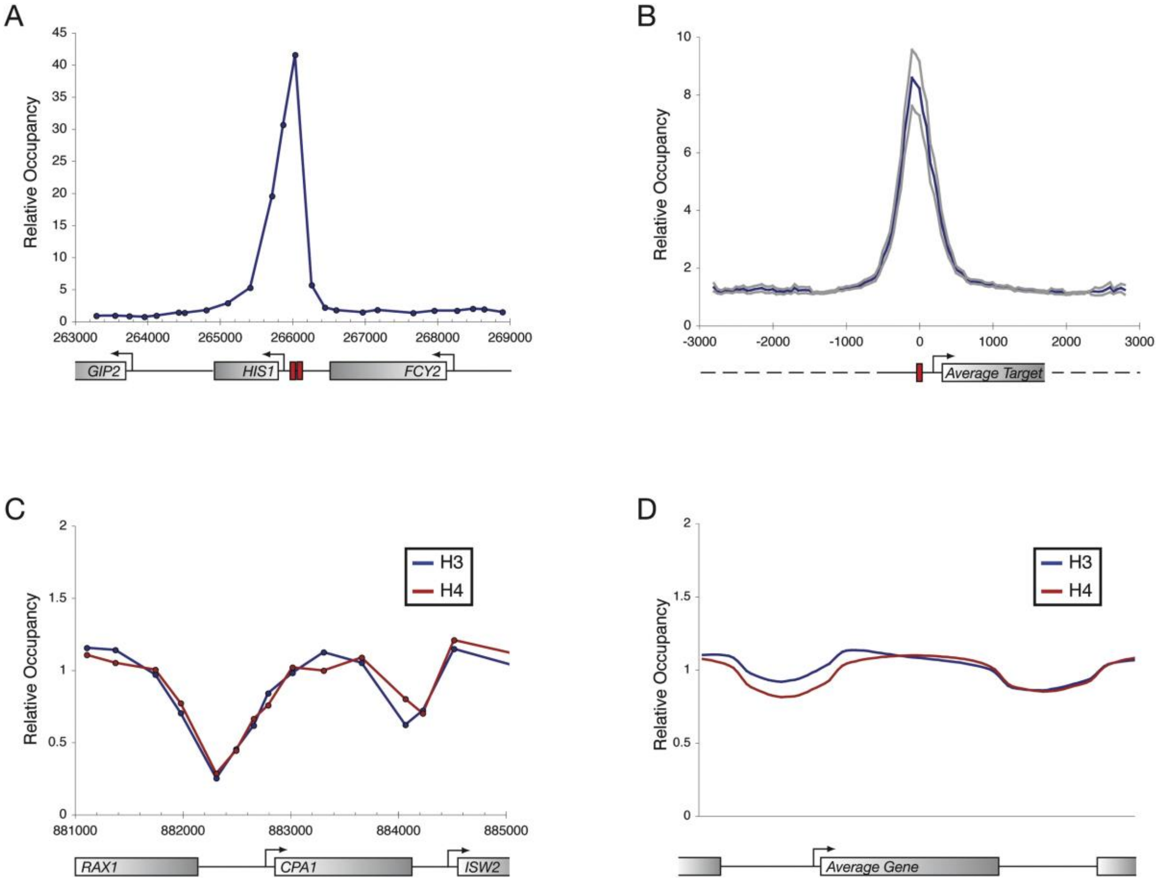


Figure 5. Occupancy of chromatin structure in regions of an ORF. A. Occupancy of transcription factors in the specified genomic region. B. Occupancy of transcription factors in the aggregated promoter region of the genome. C. Relative Occupancy of histones in the genomic region. D. Relative Occupancy of histones in an aggregated genomic region. [27]

Next Generation Sequencing as a Method to Interrogate Epigenetic Patterns

Next generation sequencing is an umbrella term for several modern methods that allow for both RNA and DNA sequencing a faster rate and cheaper than previous methods. The first human genome ever sequenced took over a decade to deliver a final product. Today this same task can be done in 24 hours. [28] Next generation sequencing has lowered the cost of research while also improving the quality of data produced. In the time since genome sequencing technology has outperformed Moore’s law in terms of cost reduction. In the year 2008 we see exponential cost reduction, and this is a result of the emergence of the second generation of sequencing technology. [29] The diminishing cost of sequencing the human genome has resulted in the market viability of genomics in healthcare and clinical applications. [30]

Deep sequencing refers to sequencing methods that provide great depth in the libraries, which is to say that the same genome is sequenced anywhere from 40 to 1000 times in a single run at a given nucleotide. The advantage of having depth is that it provides definition, coverage, and confidence to any given region of the genome. A benefit of modern sequencing techniques is multiplexing which is the capability to mix DNA samples and run them at the same time reducing the cost. This done by annealing a known sequence of DNA to the DNA fragments that will be sequence, each sample then having a unique sequence that acts as a barcode. Upon sequencing the researcher can then separate the DNA by the annealed barcode. While there are several methods and private companies that have produced whole genome sequencing methods, the dominant company is Illumina Deep Sequencing **(Figure 6)**. This is a method of sequencing that works by annealing DNA to a microarray. From here each strand is the complimented with a fluorescent nucleotide that is photographed. These photographs will chronologically be transferred from colors to the corresponding nucleotide. From these robust datasets, scientists can draw nuanced conclusions about the nature of genomic samples by mapping these samples back to the reference genome.

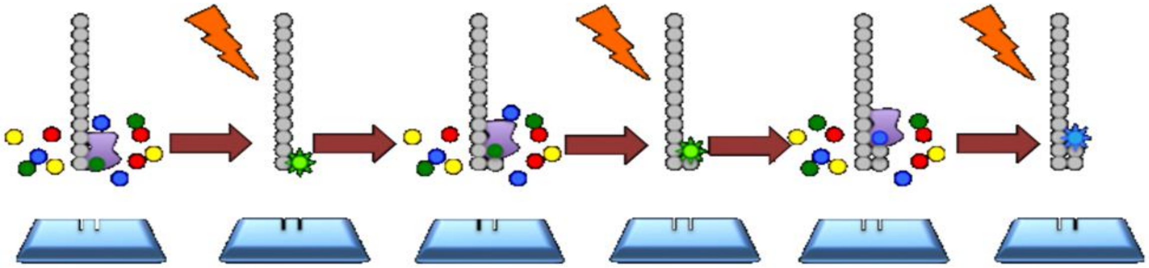


Figure 6. Process of Illumina shotgun sequencing. *[31]*Reads are annealed a slide with free nucleotide that are fluorescent. A ligase is added to add the complimentary nucleotide and a fluorescent light is used to obtain a signal giving indication as to which complimentary nucleotide was used.

A read is the portion of a DNA fragment that is sequenced. Upon receiving the results of deep sequencing, it is necessary to organize the reads as they correspond to their place in the genome to attach a genomic meaning to each sequence. One method that produces quality mapping results is paired-end sequencing. In paired-end sequencing both the 3’ and 5’ end of each read is sequenced. The first advantage of this is the higher confidence in the mapping position due to the matching algorithm being fed more precise parameters. The second advantage of paired-end sequencing is the detection of insertion and deletions. Paired-end gives the added information of DNA fragment length and in doing so gives a clearer picture of localized changes in the genome. [32]

The ENCODE project is short for encyclopedia of DNA elements. ENCODE serves as a database that seeks to annotate all regions of transcription, transcription factor association, and histone modifications. [33] In 2007 ENCODE embarked on a goal to delineate the biochemical function of every DNA element. In the time since ENCODE has created an expansive resource for scientists to further their understanding and even diagnostic ability. Upon finishing the initial project ENCODE found that 80.4% of the genome has a DNA element associated with it that could be anlayzed. For example, by mapping the binding and functionality of every transcription factor a correlation matrix can then be made. By creating an association of transcription factors we can gain insight into biological pathways. While most transcription factors interacted as expected in comparison to RNA-seq, some new correlations were made **(Figure 7)**. RNA-seq is the process of sequencing all the RNA in a sample, which gives insights to gene expression. What can be seen in the cluster **(Figure 7A)** is that these transcription factors were associated regardless of the relative position to the promoter region. In the cluster labeled B however, the association is only strong in the proximal promoter regions. The studies of histone modifications show massive variations among samples that correspond to differing levels of expression. With ENCODE the field of genetics can begin to disentangle the mechanism of disease states. [33]

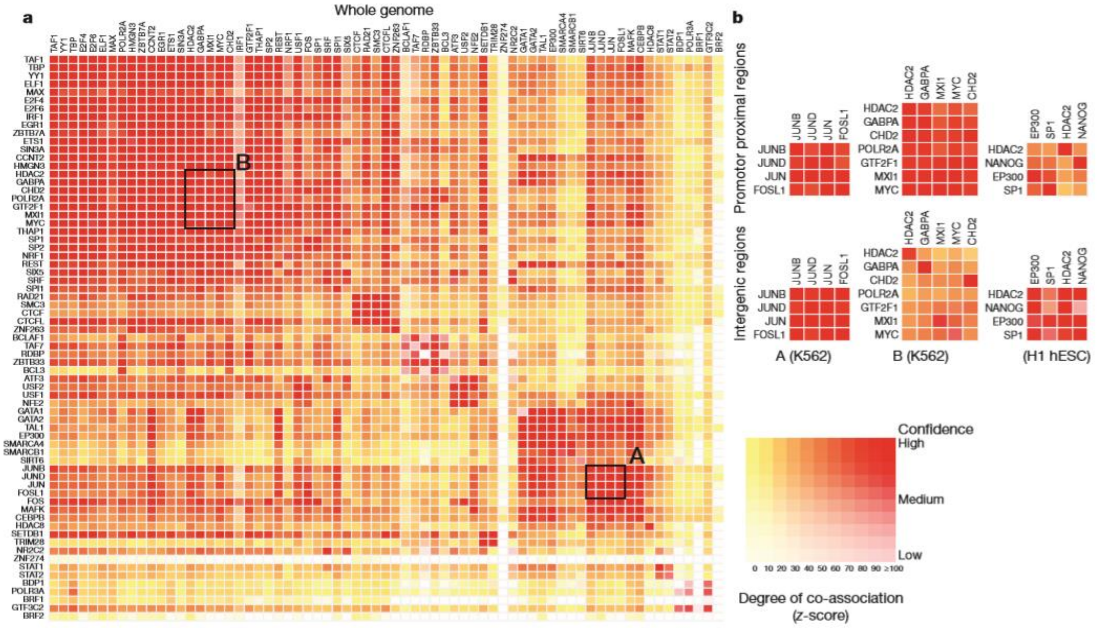


Figure 7. Transcription factor binding correlation matrix A. A correlation matrix showing hierarchical clustering was made to show Co-association between transcription factors. B. An enlargement and comparison of intergenic regions to proximal promoter regions. These regions are labeled A and B respectively on the larger matrix. [33]

Chromatin Digestion

Deep sequencing has been used to map nucleosome positioning in the past using Micrococcal Nuclease (MNase). MNase is a naturally found enzyme in bacteria that is used to assist in the digestion of viral DNA. MNase can cleave at nucleosomes and digest in chromatin pattern. [34] This has been the standard methodology in chromatin digestion, but it has two major draw backs. The first is that MNase being an enzyme is a large and may not be able to penetrate between the nucleosomes in heterochromatin. MNase complexes are the same size scale as the nucleosomes themselves creating an issue of accessibility to linker-DNA **(Figure 8)**. [35] MNase also has a pronounced bias to cleave predominately at A and T nucleotides. While it has limitation, protocols have been developed to use MNase to map the chromatin of the whole genome. [36]

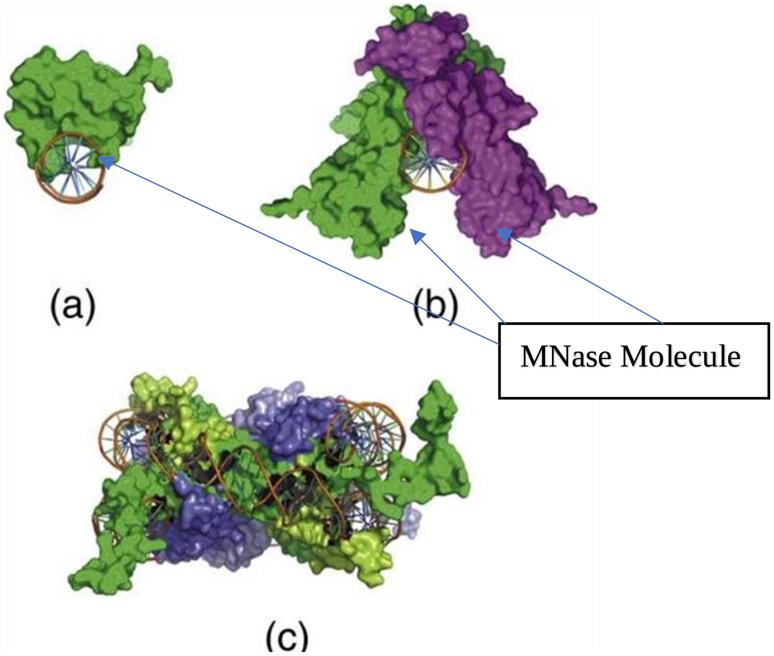


Figure 8. Space filling model of MNase bound to DNA a.) MNase cleaving DNA. b) Dimeric representation of MNase cleaving DNA. c.) Nucleosome with core particle for size reference. *[35]*

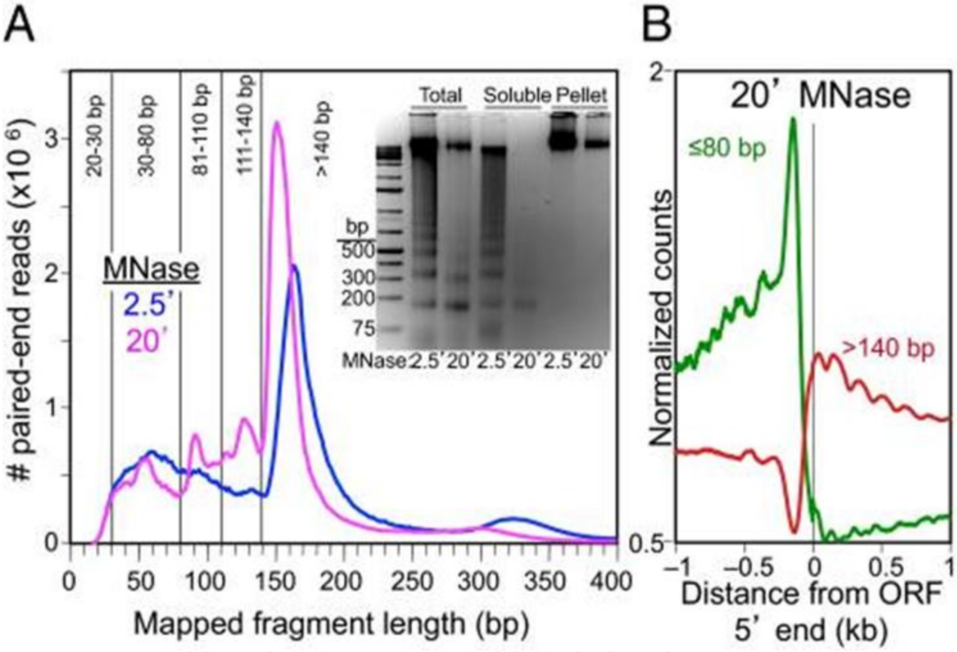


Figure 9. Read length MNase. A. Read length distribution in MNase digested libraries. B. Reads are class sized as greater that 140 base pair or less than 80 base pairs. These datasets were aggregated to an open reading frame and plotted by depth at the relative position. [36]

In 2011, the Henikoff lab published a protocol that could characterize the epigenome at single base-pair resolution. [36] The process involved the traditional use of Micrococcal nuclease in combination with paired-end sequencing. The use of paired-end sequencing allows for the researcher to draw conclusion about what particle corresponds to each epigenetic element. Different size classes show an anti-correlation when an open reading frame (ORF) is done **(Figure 9)**. It can be concluded that the base pairs smaller than 80 correspond to transcription factors since it is upstream of the gene and is comprised of smaller reads. It can also be inferred that the >140 bp size class is predominantly nucleosome bound DNA due to both the size and position relative to the start of the ORF. It is important to note the two samples studied by the Henikoff group represent two different levels of MNase digestion, 2.5 minutes and 20 minutes

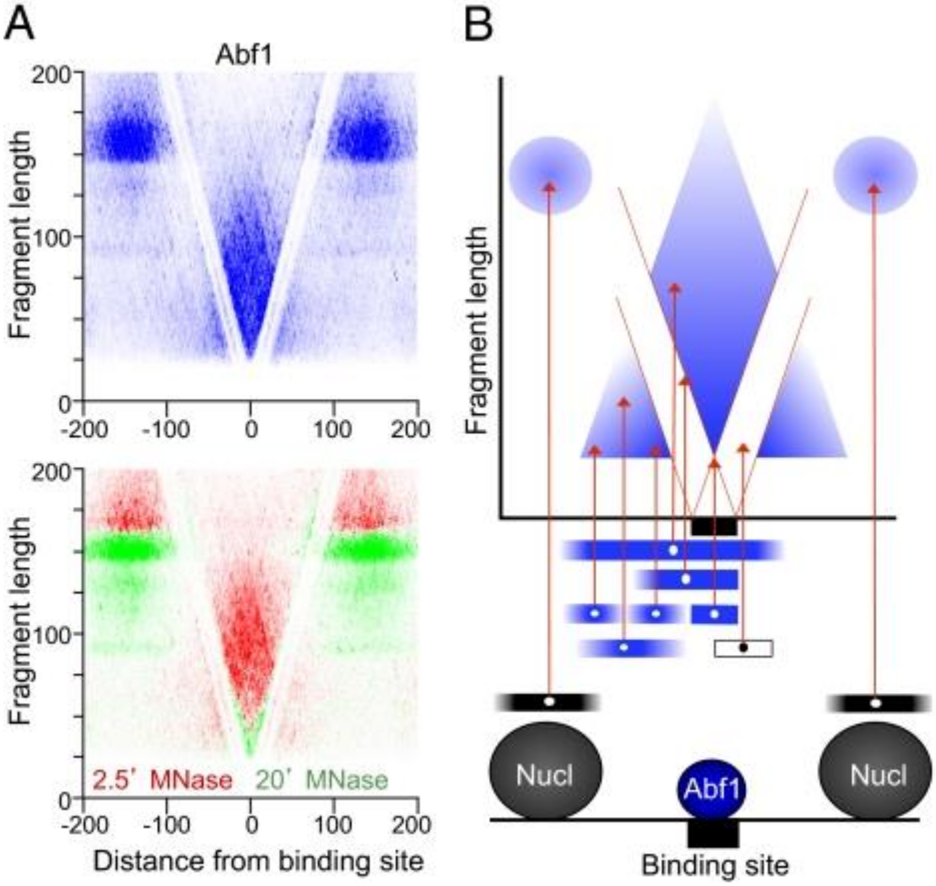


Figure 10. Aggregated V-Plot with MNase V-plot MNase chromatin digestion at the Abf1 binding site. A. an aggregation of all abf1 binding sites plotted by read length and midpoint position. [36]

(Figure 9a). In both cases, there is a large peak corresponding to nucleosomes at 150-160 base pairs as well as the peak from 300-350 likely corresponds to di-nucleosomes. In the 80-30 region there is a peak that corresponds to transcription factors. [36]

The V-plots confirm the implications made by the previous plots in that we see is that the data clusters by both genetic location and size **(Figure 10)**. The interpretation on the right side helps us to understand that the pattern of data shown corresponds to an expected epigenetic pattern of a nucleosome free region with a transcription factor in the middle. The data that corresponds to larger size-classes is clustered in the top corners of the graph and is in the location of known nucleosomes. The smaller size classes cluster at the transcriptional start site and correspond with the location of a specific transcription factor binding site. [36] The implications clearly show that paired-end deep sequencing data can be data mined to show the signatures of epigenetic data.

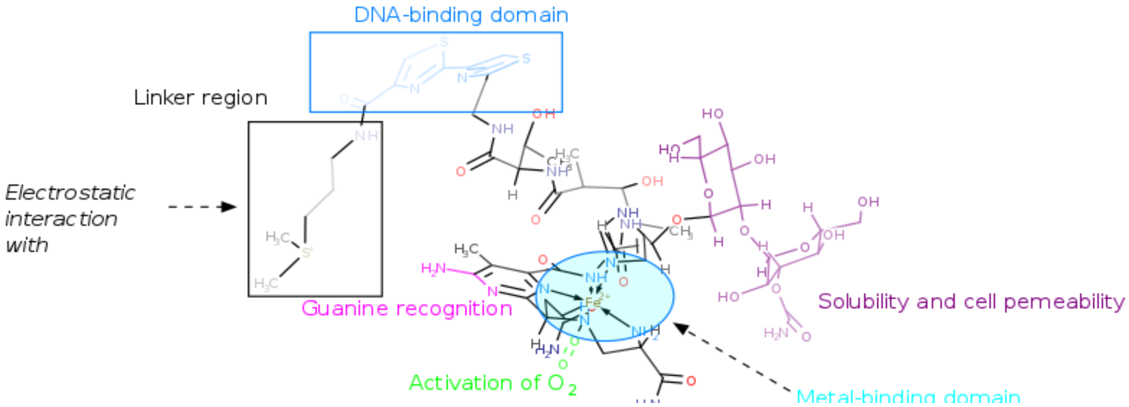


Figure 11. Molecular structure of bleomycin with labeled functional groups.[37]

The Structure of Bleomycin

Bleomycin is a chemotherapeutic drug that works by of creating a free radical that oxidizes the DNA backbone and in turn breaks the DNA strand. [38] The free radical is produced when Fe^{2+} is oxidized in the metal binding domain **(Figure 11)**. [39] Bleomycin is a chemically engineered drug, meaning that time has been taken to optimize pharmacokinetics of its structure making it ideal for cell permeability. Bleomycin is stabilized by the minor groove by hydrogen bonding with the DNA backbone as indicated by the region that shows an electrostatic interaction. Bleomycin has a molecular weight that 1.4 kDa which implies that its smaller molecular size can lead to a better digestion of heterochromatin. The typical problem with heterochromatin digestion is it is done with enzymes that are larger than the spaces between nucleosomes. Previous studies have done the preliminary molecular genetics studies to show the release of nucleosome sized fragments when Bleomycin is used as a DNA digesting agent. [40] A known way of identifying nucleosome digestive agents is the separation of DNA fragment using gel electrophoresis. When digestion was done a nucleosome ladder pattern emerged in the gel **(Figure 12)**. [41] In the gel image, lanes 1 and 2 are standardized weights, while Lanes 3 through 9 are DNA fragments that have been generated by the digestion of chromatin by Bleomycin in increasing order of increasing concentration. The nucleosome ladder is most exemplified in lanes 4 and 5 where there is the emergence of a banding pattern at mono-nucleosome, di-nucleosome, and tri-nucleosome size fragments. This confirms the ability Bleomycin to cleave at nucleosome positions and gives insight into appropriate concentrations for experimental digestion. The extent to which Bleomycin is a successful reagent for chromatin digestion has not been explored in the literature.

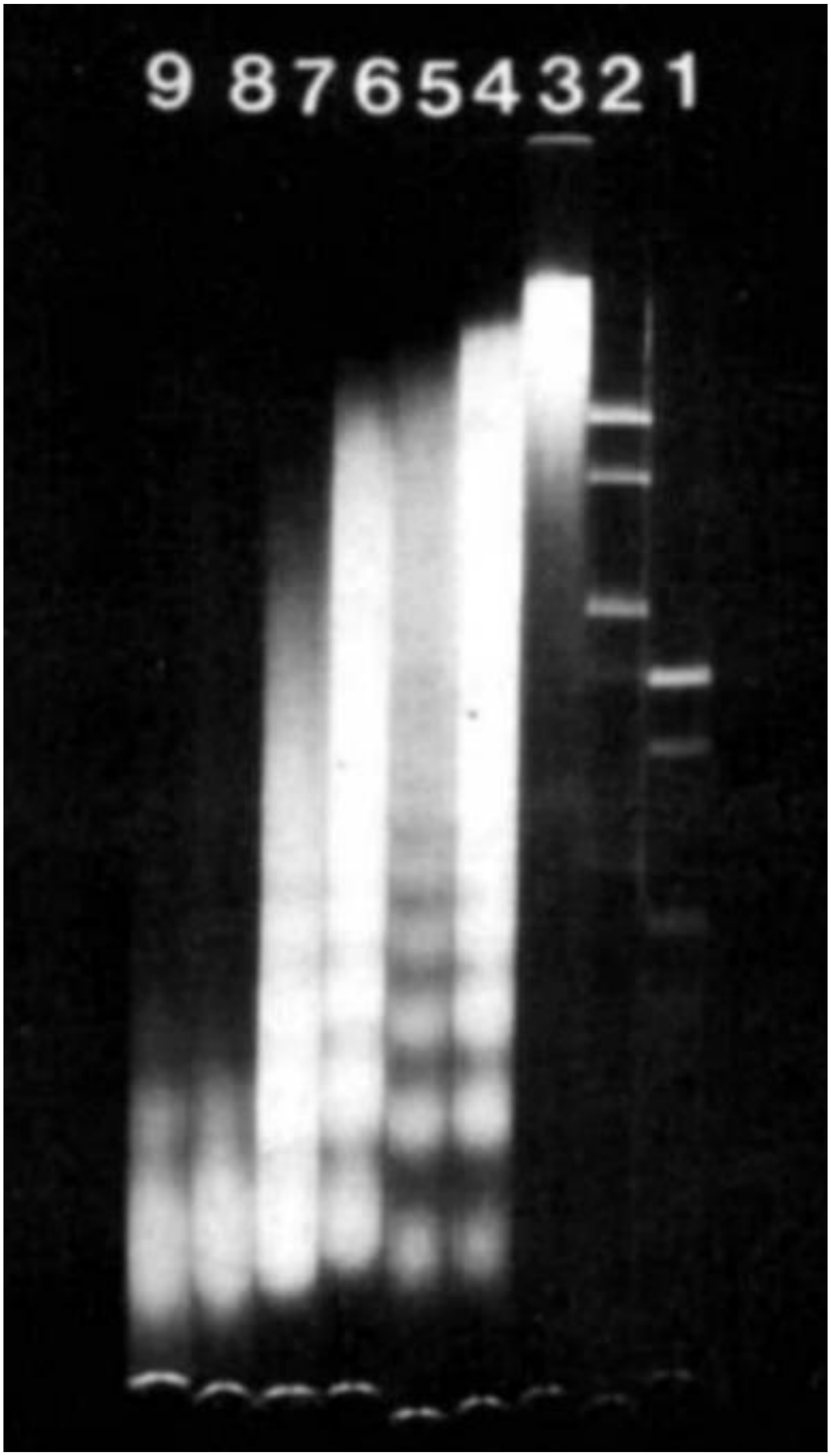


Figure 12. Initial Nucleosome banding with Bleomycin. Bleomycin digestion of DNA. Lanes 1 and 2 are weight standards. Lanes 3 through 9 are bleomycin digestion in increasing concentrations respectively. [41]

Experimental Plan

Using an established method of Next Generation Sequencing to map chromatin associated reads back to the genome, a protocol was developed to assess the viability of Bleomycin as a chromatin digestion reagent in *S. cerevisiae*. A protocol was prepared using gel electrophoresis as a means of developing libraries with mostly mono-nucleosomes. This was done at three digestion levels: 32 Units Bleomycin, 8 Units Bleomycin, 2 Units Bleomycin. The multiple digestion levels give a possibility of obtaining the footprint of smaller chromatin structures in the dataset such as transcription factors. The higher digestion libraries will theoretically contain nucleosome footprints base on the MNase data. The data analysis process will be done side by side with MNase datasets as a comparative study. This comparison will hope to give a clear display as to whether or Bleomycin can be used as an alternative to MNase. The two primary focuses in this comparison will be on nucleotide biases and nucleosome resolution in different regions of the genome. Further questions can be tested to attempt to develop novel ways of visualizing chromatin data. The profile of the data could tease out the way by which Bleomycin interacts with nucleosomes. We hypothesized that Bleomycin can be used effectively as a chromatin digestion agent to produce nucleosome positions at single nucleotide resolution.

Chapter 2

Methods

Micrococcal Nuclease Digestion and Library Preparation

Sequencing data for *S. cerevisiae* whole nuclei digested with MNase for 2.5 or 20 minutes as described in Henikoff *et al.*, 2011 was graciously provided by the Henikoff Lab.[12]

Preparation Saccharomyces cerevisiae and Bleomycin Digestion of Chromatin

Cultures of *S. cerevisiae* strains BY4741 and *PMP2-GFP* were grown in YPD media to OD-600s of 0.553 and 0.423, respectively. Four 47 mL aliquots of each culture were crosslinked using 37% formaldehyde diluted to final concentration of 1% formaldehyde. After a 15 minute incubation at 30 °C, the reaction was quenched with 2 mL 2.5 M glycine for 5 minutes at 30 °C. Each aliquot of cells was resuspended in 1 mL of bead beating buffer modified to include 1 mM EDTA. Cells were pelleted at 4000g for 3 minutes and resuspended in 1 mL of bead beating buffer with EDTA. Six µL of zymolyase [Eloo5-A] were added and the solution was incubated for 30 minutes at 30 °C. The cells were once again pelleted at 4000g for 3 minutes and resuspended in 3 mL of bleomycin digestion buffer (15 mM Tris-HCl pH 8.0, 50 µM $FeCl_3$, 5 mM DTT, and 0.5% NP-40 in dH_2O). Five hundred µL samples from each strain were treated with either 0, 0.5, 2, 8, or 32 µL of 0.1 mg/mL bleomycin for 10 minutes at 37 °C and quenched with 100 µL of 0.5 M EDTA pH 8.0 on ice. Samples were then treated with 10 µL of 20 mg/mL proteinase K and 20 µL of 10% SDS overnight at 65 °C and purified with a PCI extraction.

Gel Electrophoresis

Gels were prepared with 3% agarose in 1X TAE and visualized using ethidium bromide. When possible, samples were diluted to 500 ng/µL and run as a solution of 5 µL of DNA with 2 µL of Orange G. A 2-log ladder was used as a standard. The contrast and brightness of images were adjusted to improve visibility of bands.

Library Preparation

The PCI-purified DNA was treated with RNase A, and the ends were dephosphorylated with calf intestine phosphatase. Library preparation was followed as described in Henikoff *et al*., 2011.[36] NEB Next Ultra II DNA Library Prep Kit was used for the ligation of adaptor prior to NGS sequencing. Paired-end sequencing was performed using an Illumina HiSeq 2000 at UPenn core facilities.

Sequence Alignment, Data Cleansing

Bowtie2 was used to map the reads back to the genome using SacCer2 as a reference genome for both bleomycin and Henikoff datasets.[42] Dovetailing was used to align both sets of data. Bleomycin dataset was trimmed by 50 base pairs on the 3' end. The first 25 base pairs removed a primer that had been ligated on. The second 25 base pairs eliminated all reads under 25 bp in length for statistical accuracy. The bleomycin dataset was made in replicates and upon being fed into the bowtie2 pipeline these samples were merged into one file. The files were then converted to binary alignment format (BAM) for the preservation of computational space.[43] These files were then data cleansed by filtering for only the highest mapping scores. These files were then sorted using samtools sort by position. Each bam was then separated into 20 base pair size classes. These were then individually indexed so they could be viewed in IGV.[44]

Aggregation

Peaks were called using the DANPOS2 Dpos function.[45] Aggregation plots were generated using the ngs.plot version 2.61 ngs.plot.r function.[46] The reference genome used by ngs.plot was made using ngsplotdb and based on the sacCer2 reference genome.

Read Length Histogram

These were created using a python script on each of the original SAM files after they had been trimmed. This script function by creating a library of each read length possible and then loop through the file to create a tally at each length. These counts were then normalized to their respective libraries.

V-Plot method

A python script was made to grab the start and end position of a read in a selected slice of a SAM file. These positions were subtracted from one another to obtain the midpoint. Another python script was made to grab the read length and store it in a list. These results were then loaded into a data table in R Studio and plotted using ggplot.[47] If the plot is small enough a 2-D density plot can be applied for better visualization.

3-D Plot

The coverage tracks of a specific region in each size class were obtained by using the IGVtools count function set to 1 bp sliding window. These datasets were then merged into a matrix of their respective digestion level and reagent creating six matrixes. The matrixes could then be plotted using the surface plot function in plot.ly.

Nucleotide Representation Histograms

A python script was made that created a library of A, T, C, and G at each of the first 25 positions in the reads. This script then looped through the first 25 bp of each read and created a tally on the nucleotide present. This was run on each of the trimmed SAM files.

Codon Count Histogram

A Python script was made that created a library of every present codon at the first 3 positions in the reads. This script then looped through the first 3 bp of each read and created a tally on the nucleotide present. This was run on each of the trimmed SAM files.

Correlation Matrix

The coverage tracks of a specific region in each size class were obtained by using the IGVtools count function set to 1 bp sliding window. These data were then put into a single data table that was transformed into a Euclidean distance matrix. This was then plotted as a correlation matrix that was hierarchical clustered.

Coverage Track for Whole Read

A 20 base pair sliding window was used in IGV count to output a .wig file that created an accurate representation of the coverage track.

Chapter 3

Results

Digestion of Chromatin using Bleomycin

To successfully test the hypothesis that Bleomycin can digest chromatin, an experimental protocol had to be initially developed. The first step in this process was to determine the appropriate concentration of Bleomycin to be used. Conceptually, the goal is to have a high digestion that resolves single nucleotide fragments and a low digestion that retains sub-nucleosome sized particles such as transcription factors. A good way to resolve this is to digest chromatin and run the DNA fragments on a gel **(Figure 13)**. This experiment was run on both BY4741 and PMP2-GFP strains of yeast. As digestion increases we see the emergence of a nucleosome banding pattern. At the highest digestion level of 32 mg/ml Bleomycin we see the darkest band at a mono-nucleosome fragment length. For this reason, 2 mg/ml, 8 mg/ml, and 32 mg/ml were selected as a range of digestion levels. In alignment with MNase digestion library preparation, the Solexa preparation protocol was modified to preserve shorter fragments.

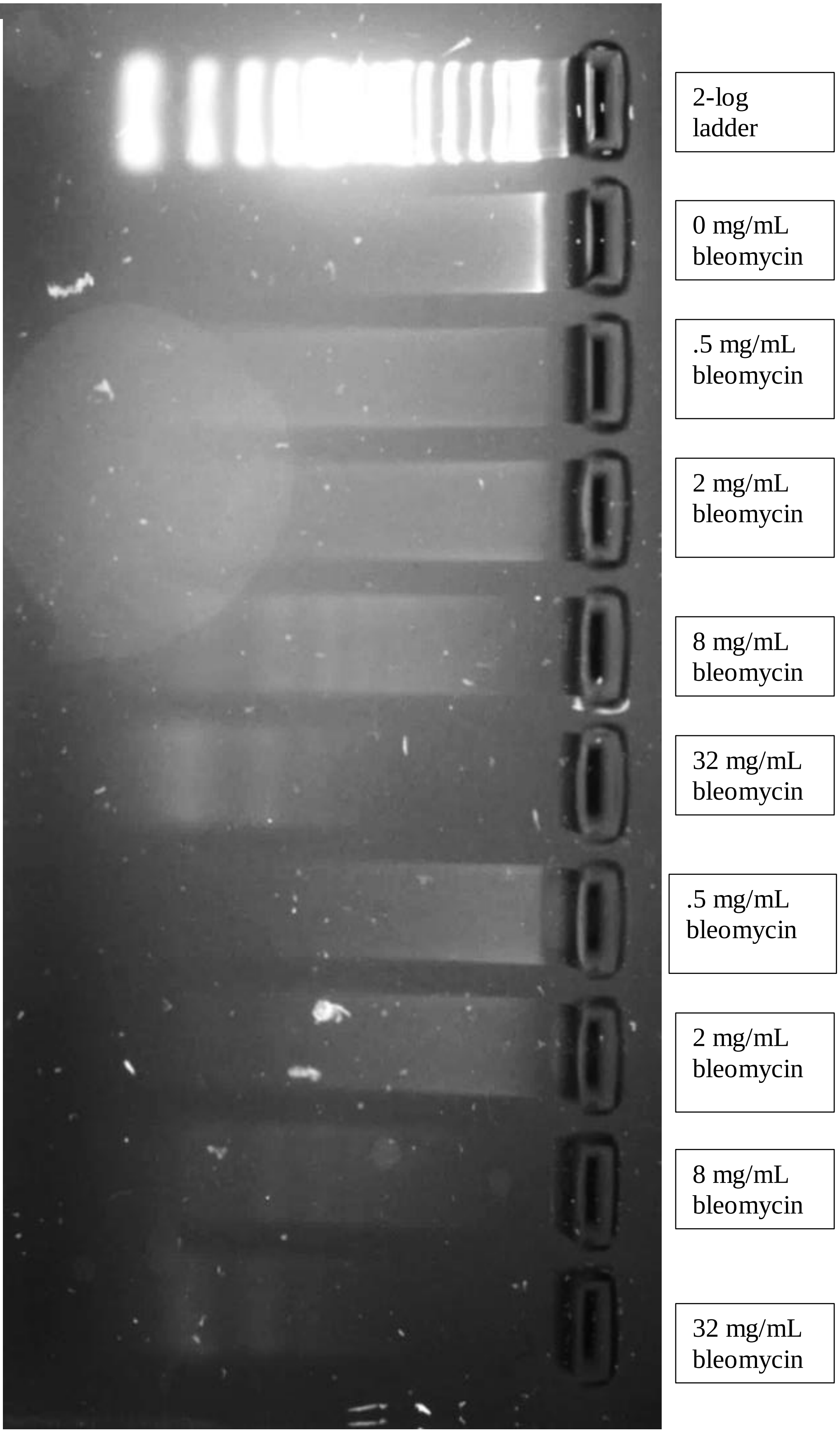


Figure 13. Banding pattern of S. cerevisiae genomic DNA following bleomycin digestion with zymolyase. Samples from strain BY4741 are in lanes 2-6, and samples from strain PMP2-GFP are in lanes 6-10.

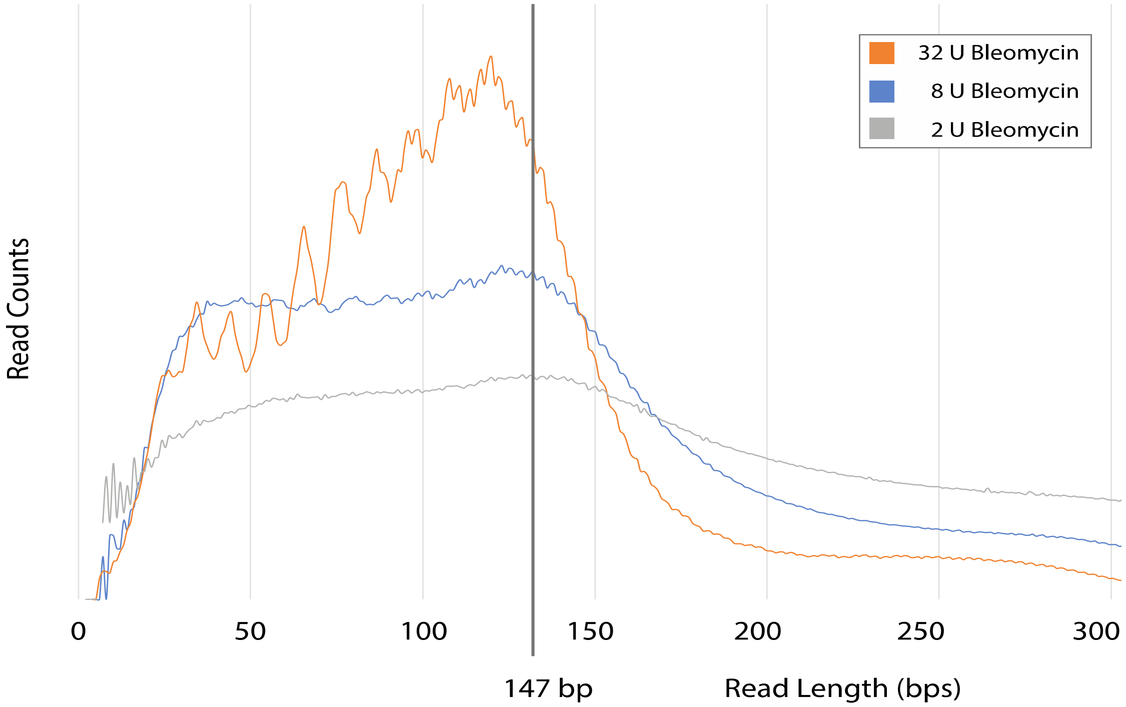


Figure 14. Normalized fragment length distributions for Bleomycin digestion libraries.

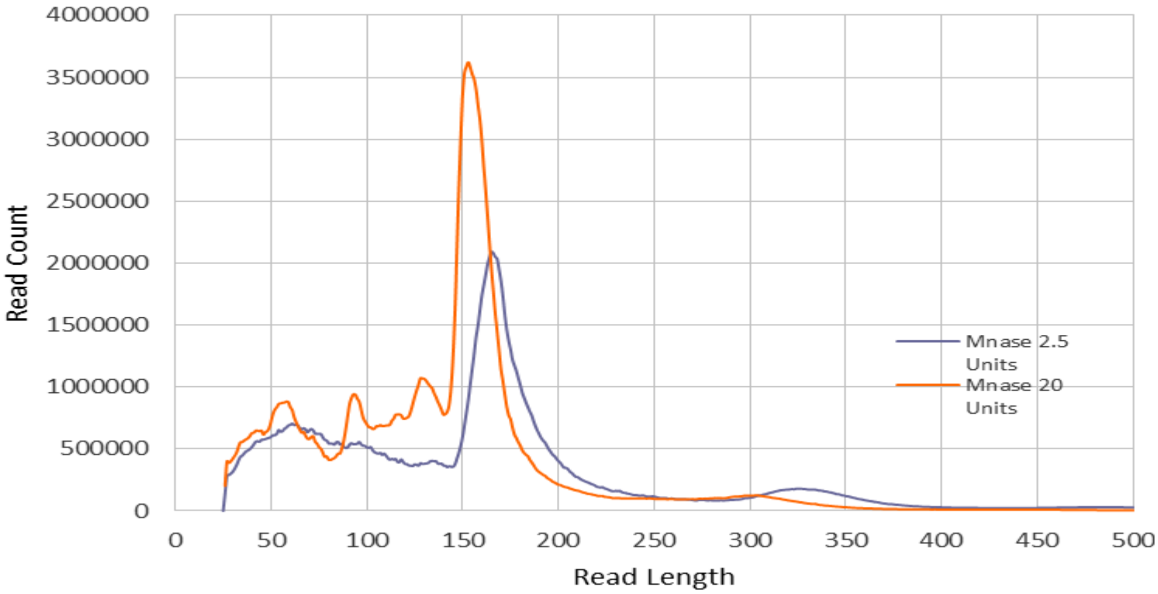


Figure 15. Normalized fragment length distributions for MNase digestion libraries.

One of the first ways the fragments in these libraries can be analyzed is by creating a histogram of fragment sizes. Both digestive agents produce read fragments length in ranges that correspond to mono-nucleosomes and transcription factors **(Figures 14&15)**. In the dataset from the Henikoff group's 2011 paper, the nucleosome peak is at 160 nucleotides in high digestion and 180 nucleotides in low digestion. MNase 2.5 Units appears to have a di-nucleosome peak at 325 nucleotides. In Bleomycin libraries the nucleosome peak is 147 bp precisely. These libraries have an emergence of a 10 bp periodicity that increases as digestion level increases **(Figure 15)**. There's a peak at 25 bp that represents transcription factors. These libraries are normalized to themselves and are robust libraries with high read counts so this is not likely to be statistical noise.

Pattern of Distinct Particles in Intergenic and Transcriptional Start Sites

To further analyze the distribution of these DNA fragments across the genome we 120 bp, 121-140 bp, and 141-160 bp. To compare the ability of MNase and Bleomycin to map transcription factors in the genome, aggregation plots were made centered on transcriptional start sites (TSS). These were binned by read size to separate nucleosome and transcription factor signatures. Both methods showed similar profiles in this analysis. The shorter size fragments (21-100) show signatures most like that of transcription factors **(Figure 16)**. This can be shown in the enriched values in nucleosome free regions (NFR) and the depletion of those values in the NFR +1 position. Respectively, the nucleosome signature was present in large fragment sizes (101-180). This signature was also depleted in the NFR region and enriched at the NFR +1 position.

To compare the ability of MNase and Bleomycin to position nucleosomes reads an aggregation plot was made where the datasets were centered on a nucleosome. Both datasets performed similarly (**Figure 17 & 18**). We see a depletion in the center position of small reads (21-60). In larger read lengths (101-180) there is an enrichment at the centered position with an equally spaced pattern of depletion followed by enrichment **(Figure 17)**. This is explained by nucleosomes be equally spaced by linker-DNA. This pattern has less definition in the largest datasets in Bleomycin and this could be due to over digestion.

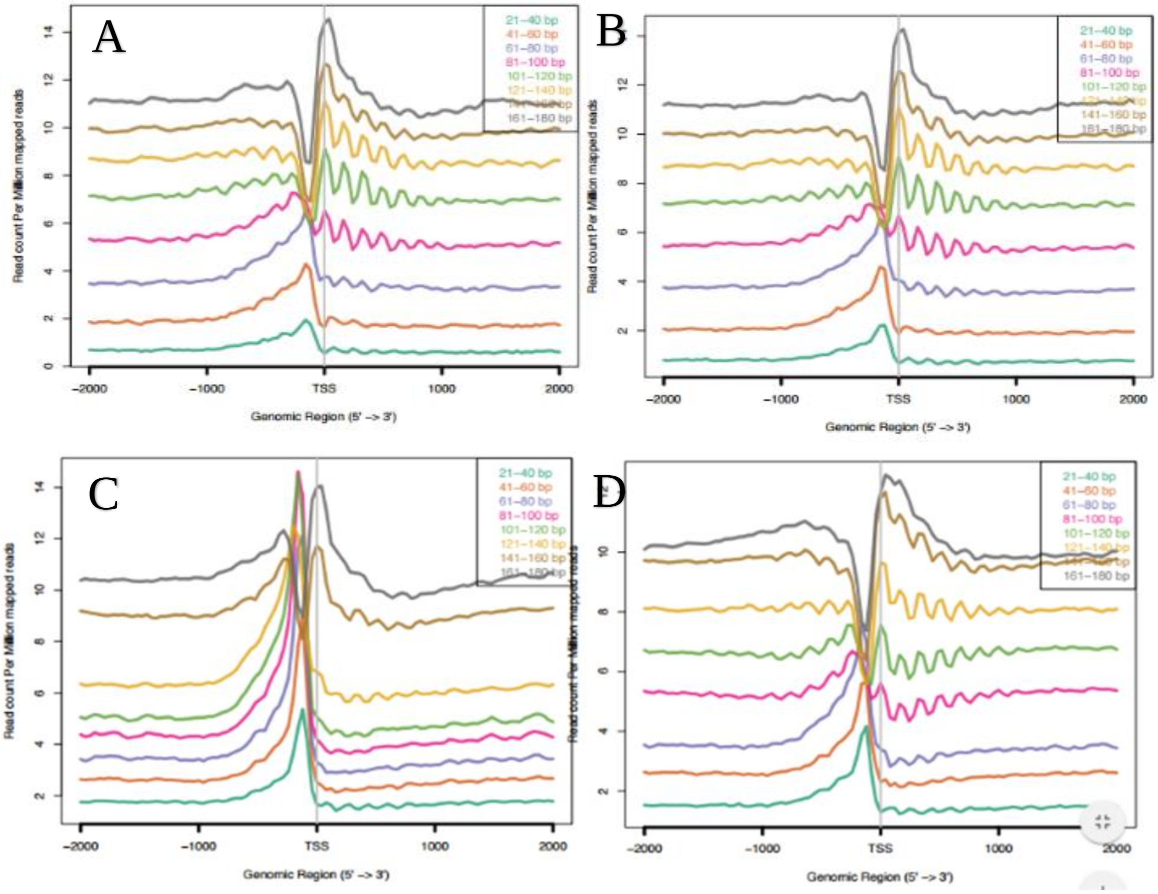


*Figure 16. Rea*d aggregation plots centered around TSSs. A. bleomycin 15 minutes digestion B. bleomycin 30 minute digestion C. MNase 2.5 unit digestion D. MNase 20 unit digestion. These plots display the overlapping of every TSS in genome and the relative coverage at that position

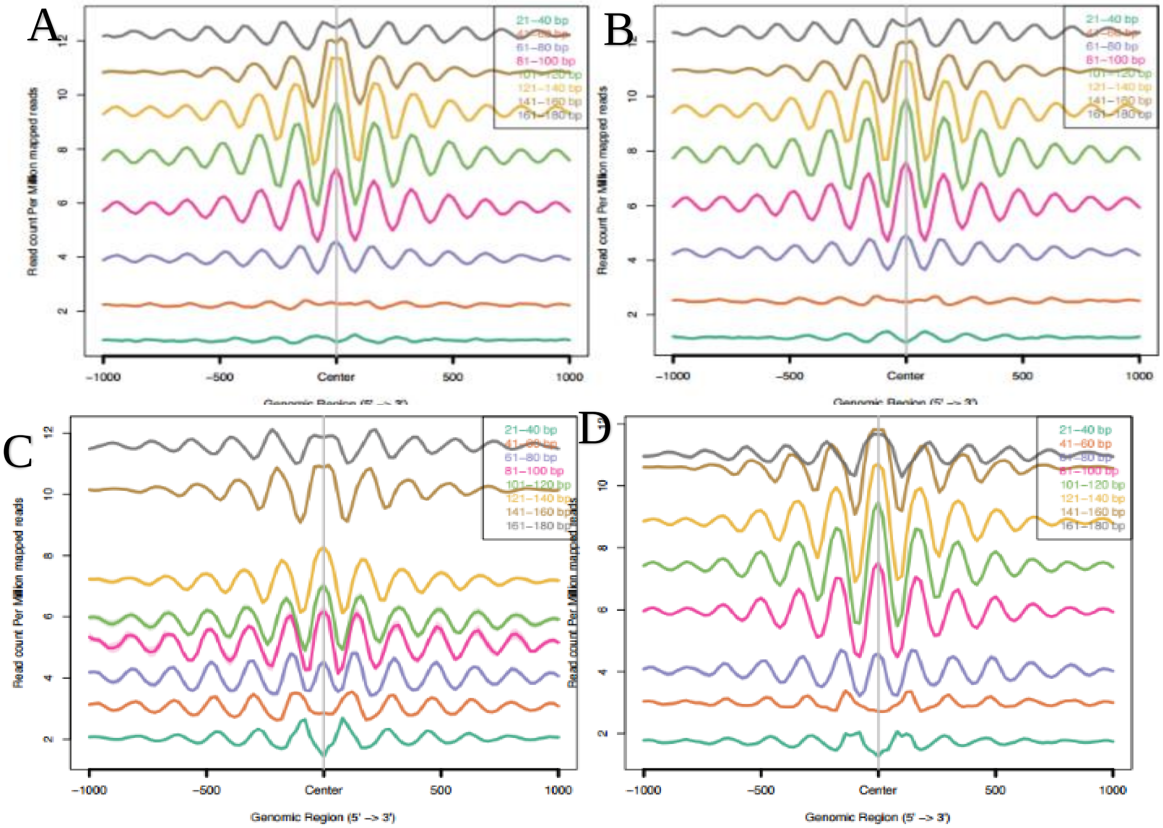


Figure 17. Read aggregation plots centered around the locations of nucleosomes as determined by DANPOS2. A. MNase 2.5 units digest B. MNase 20 units digest *C.* bleomycin 15 minute digest D. bleomycin 30 minute digest. These plots are the overlap of every nucleosome peak and the coverage in the data at the relative positions. The patterns for each size class are comparable across all conditions.

Reads can be mapped back to the genome and visualized using a guided user interface call IGV (**Figure 18)** As can be seen by the reads track, the sequencing was done on paired-ends and has to be mapped as such. An important note is that the automated coverage track, second from the top in Figure 18, only the ends of the reads are counted and as result create an inaccurate coverage track. The above track is the adjusted to be an accurate reflection of coverage and can be used to visualize chromatin structures.

Figure 19 is a display of IGV tracks on the same region shown in **Figure 23 and 24**. The first thing to notice is that the coverage tracks are coverage for the pair ends and not the entire read. Viewing the coverage track this way give the ability to see visualize points of digestion by the reagents. As previously referred to in the 436000-437000 base pair position region, we see a lack of resolution in the MNase digest. This lower resolution can be seen to be caused by the reads track having a large amount of longer reads that are overlapping. The over-digestion by bleomycin allows separation between nucleosome positions and can be the reason for higher resolution in closely positioned nucleosomes.

The coverage of the GAL1 and GAL10 gene can be visualized with our datasets for closer study (**Figure 20**). This is interesting because a remodeling of the structure of chromatin (RSC) is known to be in this region. This by an abnormally shaped structure and two separate structures present within different size classes indicated by the yellow bars. This corresponds to one chromatin structure being present in some cells and a different chromatin structure being present in the rest. We see a square peak in the GALl4 upstream activation sequence region of larger size classes of the MNase dataset (**Figure 20**). In smaller size classes in the same position there is a small chromatin fragment that maybe a transcription factor. This same position is not as clear in the bleomycin dataset there is consistent coverage over multiple size classes at the position resulting in the suggestion of chromatin remodeling. [48]

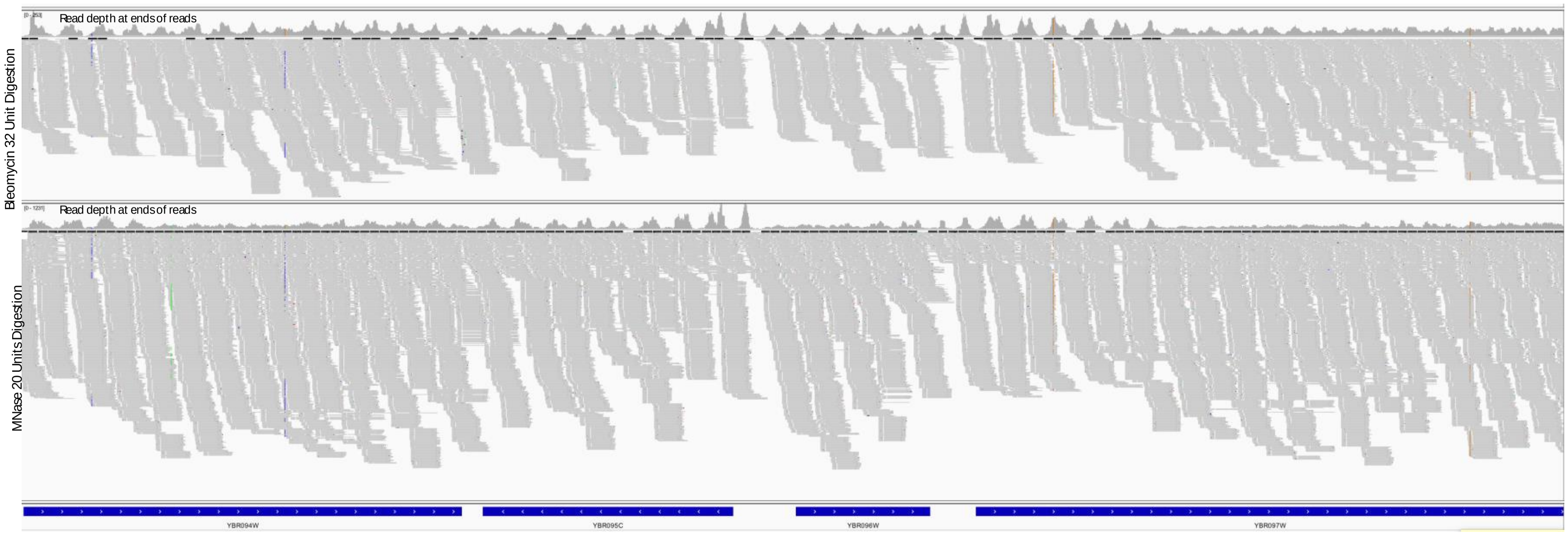

Bleomycin 32 Unit Digestion
[0 - 253]
Read depth at ends of reads
MNase 20 Units Digestion
[0 - 1231]
Read depth at ends of reads
YBR094W
YBR095C
YBR096W
YBR097W

Figure 18. Reads mapped back to the genome in high digestion of bleomycin and MNase.This is an IGV track display showing coverage and reads for 32 Units Bleomycin and 20 Units MNase.

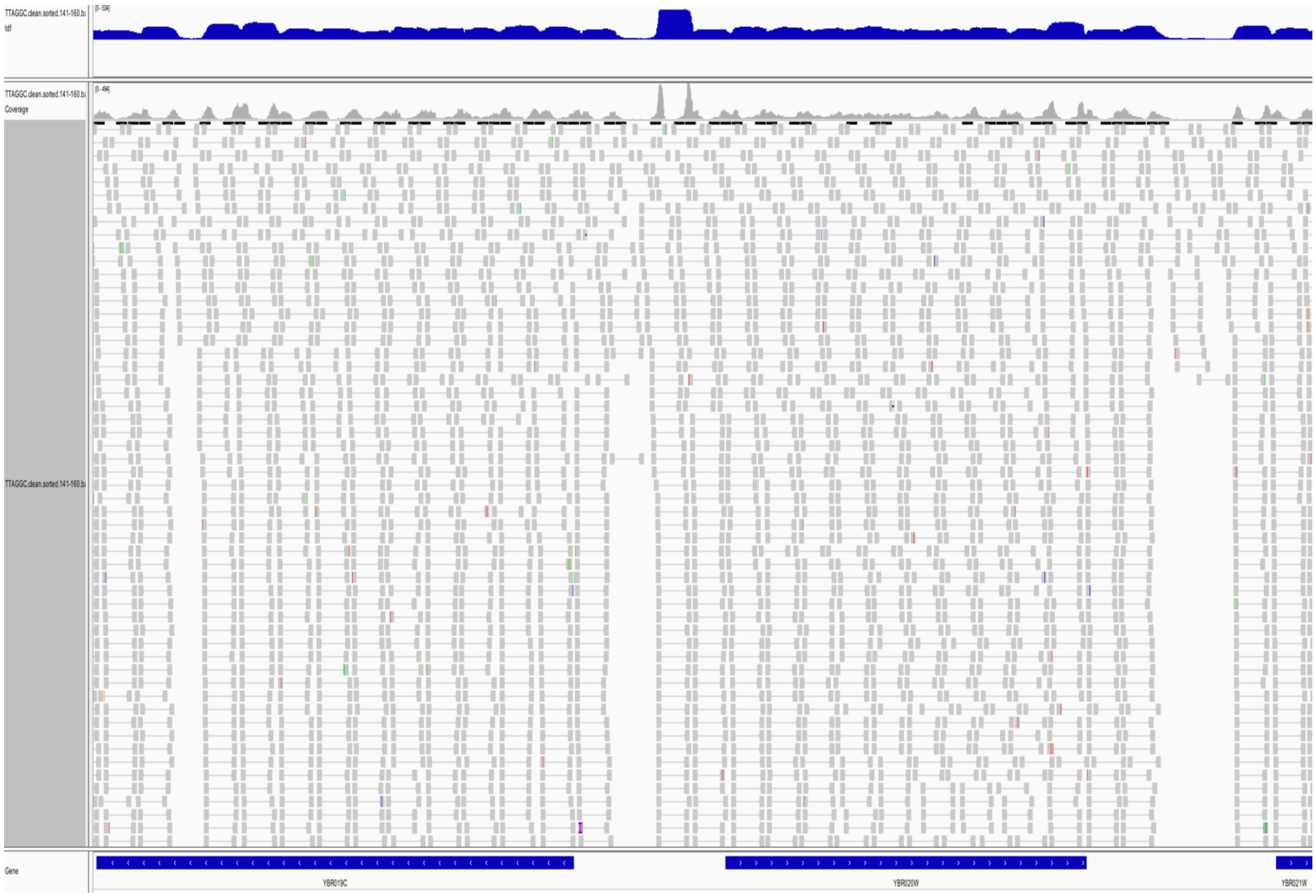

TTAGGC.clean.sorted.141-160.ba
tdf
[0 - 534]
TTAGGC.clean.sorted.141-160.ba
Coverage
[0 - 494]
TTAGGC.clean.sorted.141-160.ba
Gene
YBR019C
YBR020W
YBR021W

Figure 19. Read Mapping in IGV 32 Units Bleomycin. Tracks from top to bottom: Coverage track counting the depth of the entire read, coverage track counting the read depth for only the paired-ends, mapped paired reads back to the reference genome, gene track showing open reading frames and the direction of transcription.

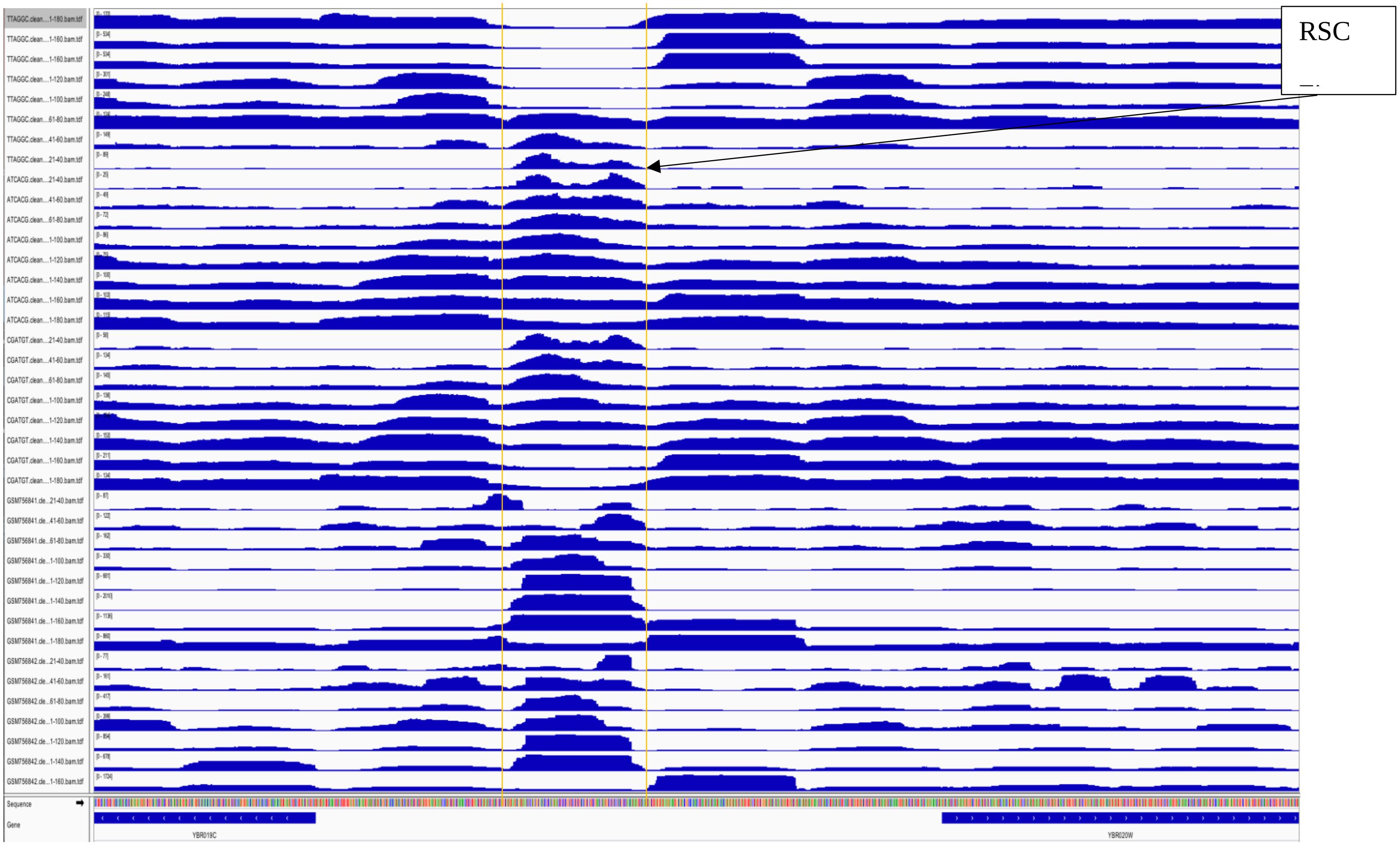

RSC
TTAGGC.clean....1-180.bam.tdf
TTAGGC.clean....1-160.bam.tdf
TTAGGC.clean....1-160.bam.tdf
TTAGGC.clean....1-120.bam.tdf
TTAGGC.clean....1-100.bam.tdf
TTAGGC.clean....61-80.bam.tdf
TTAGGC.clean....41-60.bam.tdf
TTAGGC.clean....21-40.bam.tdf
ATCACG.clean....21-40.bam.tdf
ATCACG.clean....41-60.bam.tdf
ATCACG.clean....61-80.bam.tdf
ATCACG.clean....1-100.bam.tdf
ATCACG.clean....1-120.bam.tdf
ATCACG.clean....1-140.bam.tdf
ATCACG.clean....1-160.bam.tdf
ATCACG.clean....1-180.bam.tdf
CGATGT.clean....21-40.bam.tdf
CGATGT.clean....41-60.bam.tdf
CGATGT.clean....61-80.bam.tdf
CGATGT.clean....1-100.bam.tdf
CGATGT.clean....1-120.bam.tdf
CGATGT.clean....1-140.bam.tdf
CGATGT.clean....1-160.bam.tdf
CGATGT.clean....1-180.bam.tdf
GSM756841.de...21-40.bam.tdf
GSM756841.de...41-60.bam.tdf
GSM756841.de...61-80.bam.tdf
GSM756841.de...1-100.bam.tdf
GSM756841.de...1-120.bam.tdf
GSM756841.de...1-140.bam.tdf
GSM756841.de...1-160.bam.tdf
GSM756841.de...1-180.bam.tdf
GSM756842.de...21-40.bam.tdf
GSM756842.de...41-60.bam.tdf
GSM756842.de...61-80.bam.tdf
GSM756842.de...1-100.bam.tdf
GSM756842.de...1-120.bam.tdf
GSM756842.de...1-140.bam.tdf
GSM756842.de...1-160.bam.tdf
Sequence
Gene
YBR019C
YBR020W

Figure 20. Coverage tracks between Gal10 and Gal1 genes showing chromatin remodeling complex.Top portion of the IGV window displays coverage tracks for the entire read. These are size separated files that correspond to different digestions. The digestion levels are labeled as follows 20 Units MNASE = GSM75842, MNASE 2.5 = GSM75842, 32 Units BLM = TTAGGC, 8 Units BLM =CGATGT, 2 Units BLM = ATCACG.

Fragment Midpoint vs. Fragment Length of Nucleosome and Sub-nucleosome particles

Aggregation and occupancy maps only consider length on a gross level and cannot achieve a high resolution in the data from individual fragments. As a solution to this V-plots have previously been used to better resolve the relationship between size and mid-point position (**Figure 21**). This plot is defined by plotting the midpoint position of a read against the read length. The advantage of this is that different chromatin structures will cluster at different positions and expected lengths. In this case we are expecting nucleosomes to be 147 nucleotides and transcription factors to be from 25-60 nucleotides. Reads align to the position of chromatin (**Figure 21 B**).

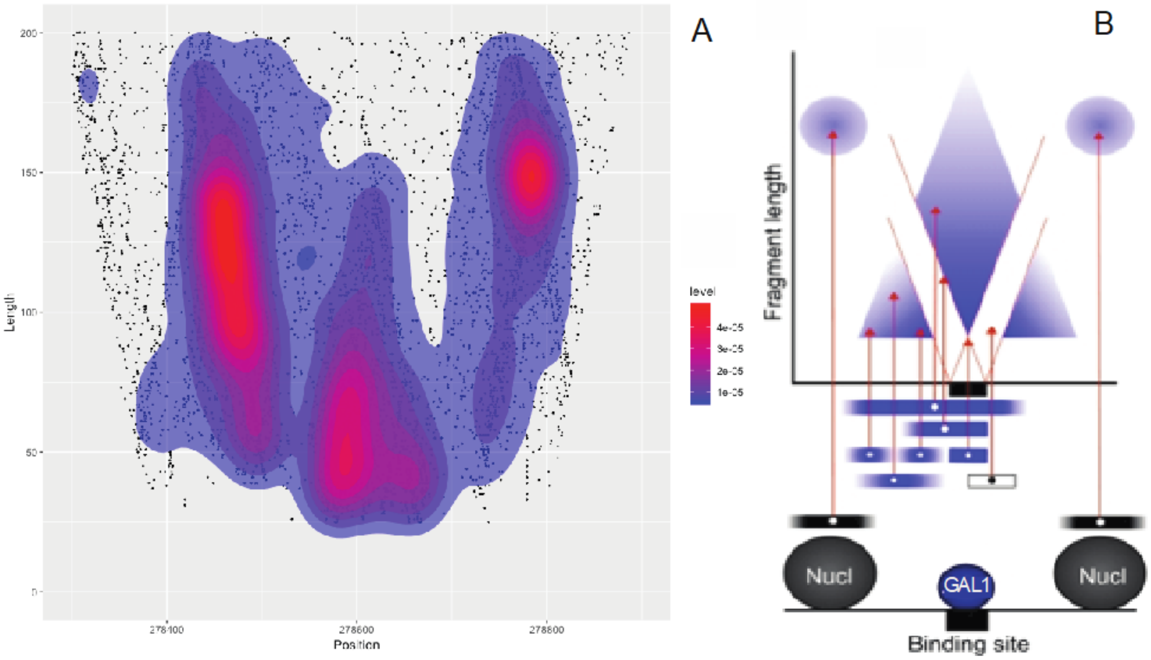


Figure 21. 2-D Density V-plot A.) V-plot of GAL4 transcription start site (TSS) showing midpoint read position plotted by length. A 2-D density map using ggplot2 was created showing the highest concentrations in red and the lowest in blue. B.) a guide to reading V-plots showing that nucleosomes are in the larger read lengths and the smaller reads correspond to transcription factors. An area between these clusters is cleared creating a “V’ shape. This corresponds to where digestion occurred. [36]

Specifically, the data when aligned to a TSS will form a “V”. This clearing in the figure visualizes regions of the genome that were exposed and digested by the reagent (**Figure 21A**). We can see that the center point is the GAL4 transcription factor binding which corresponds to the cluster in data around 25 to 60 base pairs. In this case the GAL4 transcription factor is just upstream of the GAL1 gene and it can be assumed to be expressed or poised given the positioning of chromatin and transcription factor. The downstream nucleosome is much better positioned, and this can be seen by the fact that it has far less area in its cluster. The upstream nucleosome can be said to be “fragile” as the data appears in a smeared fashion. This indicated that its position has far more variation in the cell culture resulting in greater variation in the read positioning.

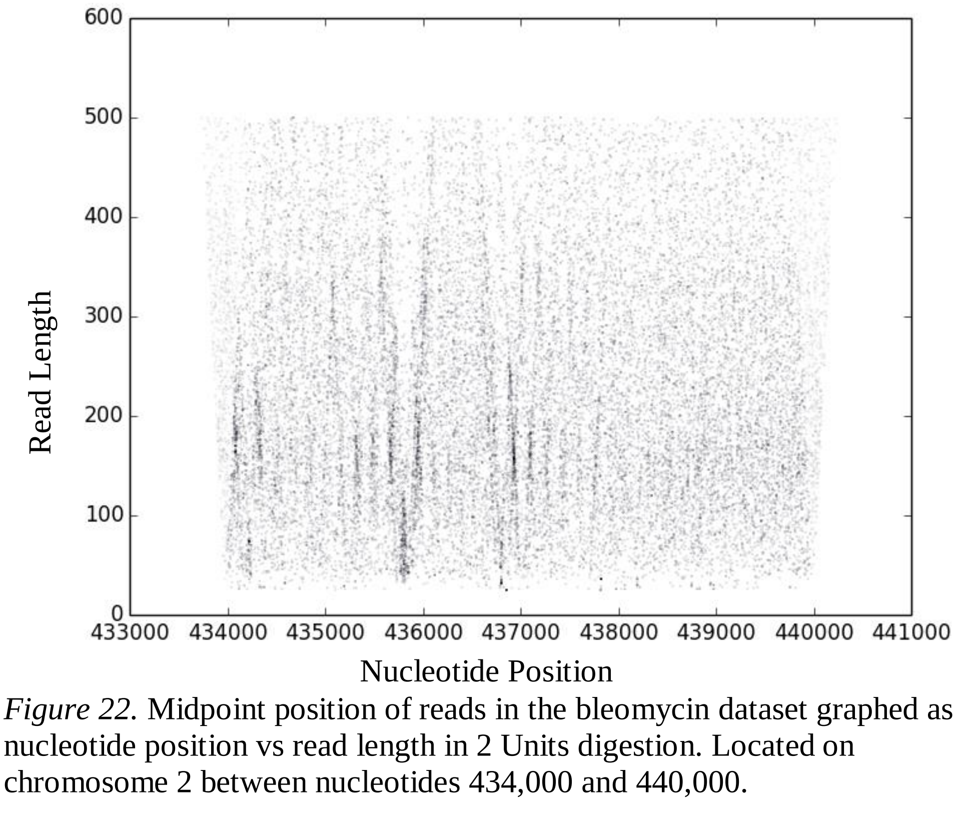


Figure 22. Midpoint position of reads in the bleomycin dataset graphed as nucleotide position vs read length in 2 Units digestion. Located on chromosome 2 between nucleotides 434,000 and 440,000.

Figures 22, 23, and 24 are the application of the V-plot method across a 6,000 base-pair region of the genome. What is evident is that different digestion levels provide different quality of data for specific chromatin structures when bleomycin is use. In this case we see that 32 Units digestion can triangulate nucleosome position down to a specific point in the genome and resolves closely positioned nucleosomes seen by the clusters around 147 bp in length (**Figure 24**). This digestion however fails to create a transcription factor footprint due to over digestion. The transcription factor footprint is clearly seen in just before the 436000 positions reflecting that lower digestion levels have the capacity to capture data signature for this chromatin structure (**Figure 22**).

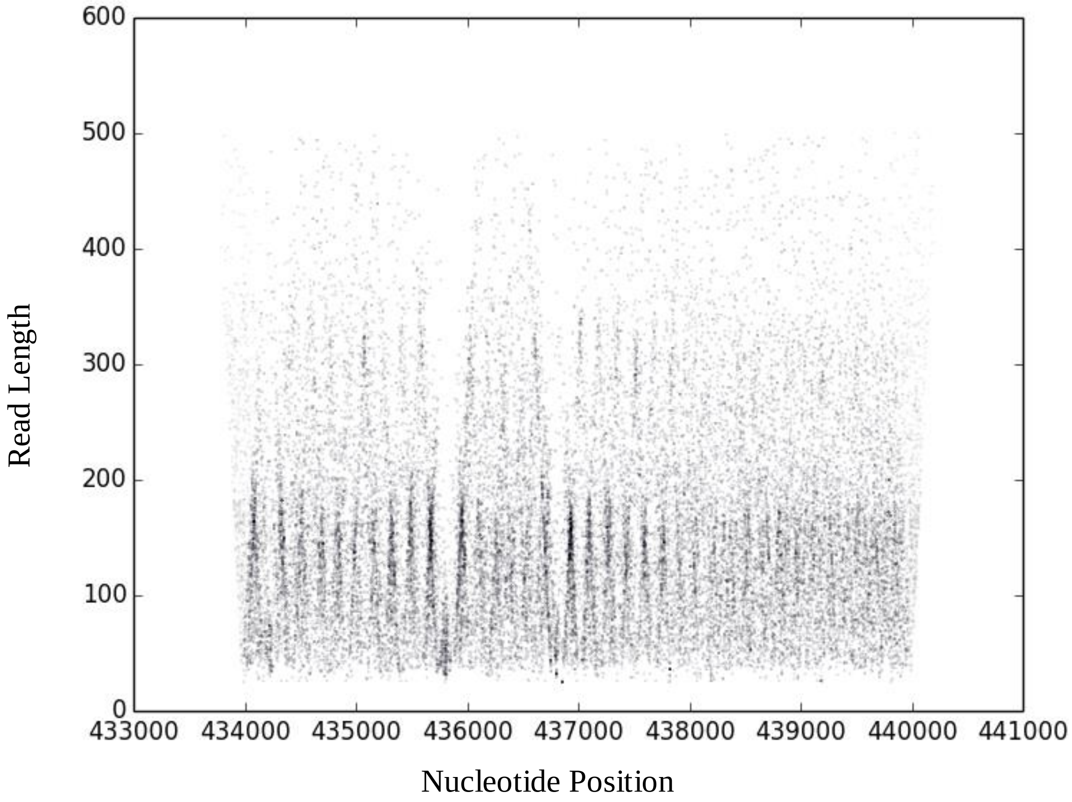


Figure 23. Midpoint position of reads in the bleomycin dataset graphed as nucleotide position vs read length in 8 units digestion. Located on chromosome 2 between nucleotides 434,000 and 440,000.

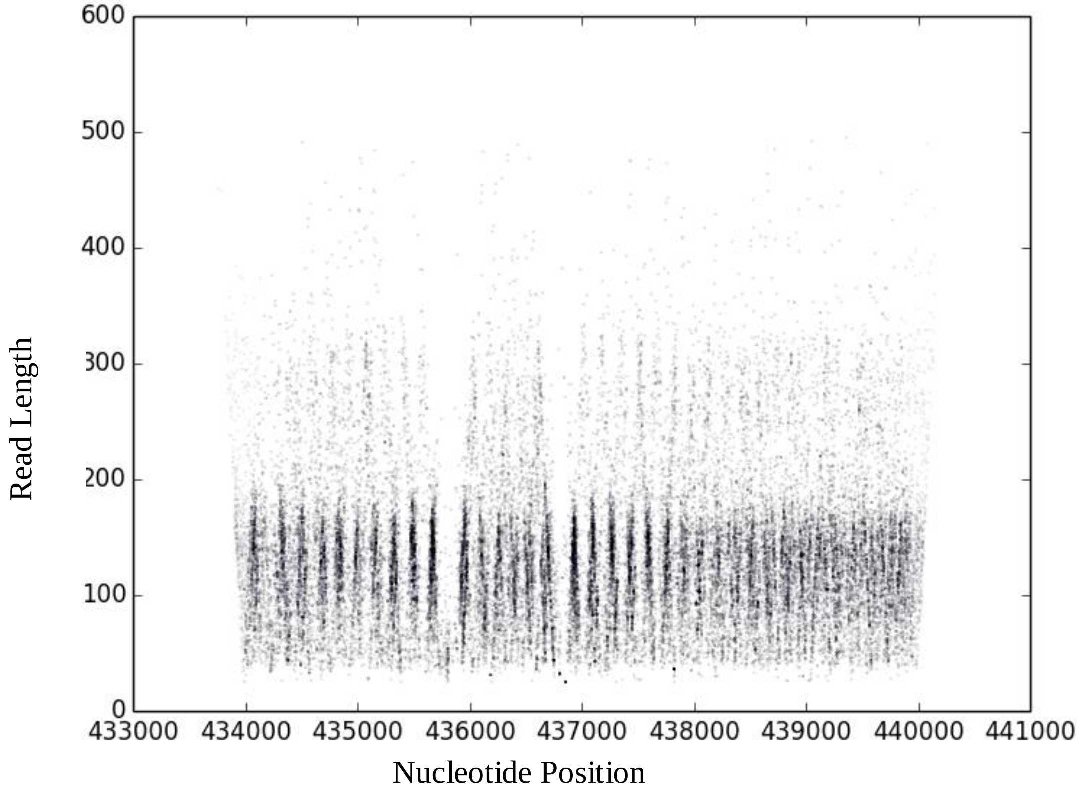


Figure 24. Midpoint position of reads in the bleomycin dataset graphed as nucleotide position vs read length in 32 Units digestion. Located on chromosome 2 between nucleotides 434,000 and 440,000.

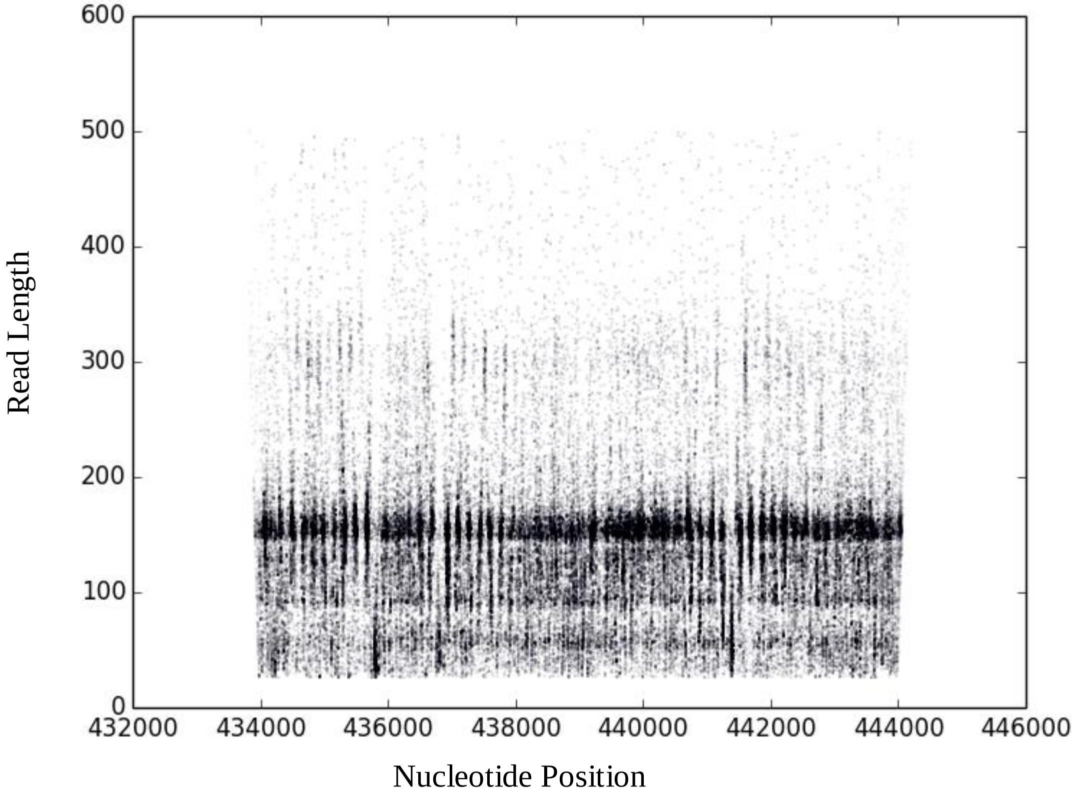


Figure 25. Midpoint position of reads in the MNase dataset graphed as nucleotide position vs read length at 20 units digestion. Located on chromosome 2 between nucleotides 434,000 and 440,000.

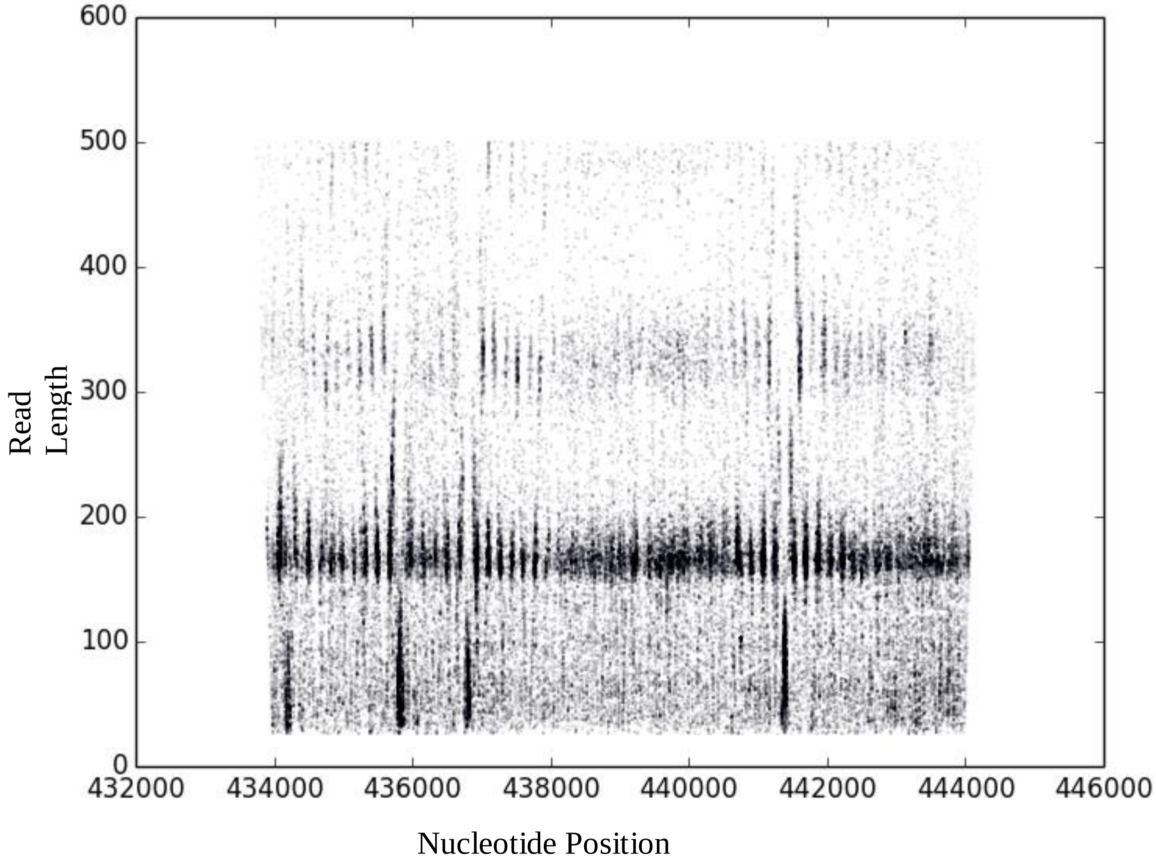


Figure 26. Midpoint position of reads in the MNase dataset graphed as nucleotide position vs read length at 2.5 units digestion. Located on chromosome 2 between nucleotides 434,000 and 440,000.

Figure 25 visualizes a major problem in MNase Digestion. The reads are oversized and as a result even in high digestion they cannot triangulate because of the overlapping of linker-DNA seen by dark band at 160 bp. We see this most in the nucleosomes around the 438000 mark in comparison to the same region in **Figure 22** is much lower resolution. This digestion also over digested the transcription factor footprint as seem by the limited clustering in 25-60 base pair in the TSS region. 2.5 Units digestion MNase shows a significant footprint for transcription factor signatures in the TSS showing (**Figure 26**). The well positioned nucleosomes are moderately resolved following and before the TSS however again at the 438000 region there is no resolution at all. Another interesting signature of this data are the clusters between 300 and 400 base pairs in length that are positioned between nucleosomes. This likely is a representation of dinucleosomes left from under digested chromatin.

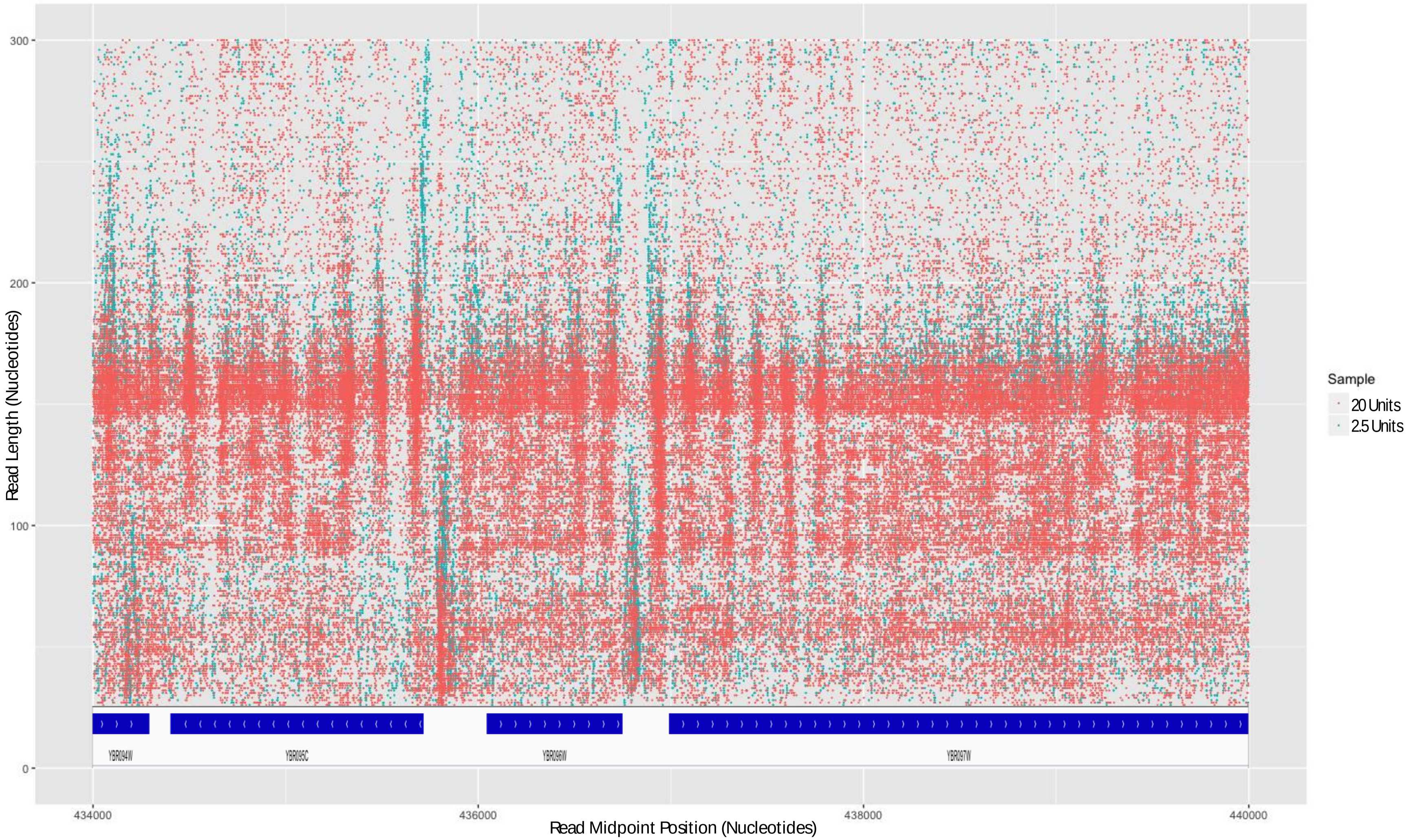


Figure 27. Midpoint position of reads in the MNase dataset graphed as nucleotide position vs read length. This is located on chromosome 2 between nucleotides 434,000 and 440,000. This is overlaid with a track showing gene position, name and orientation below it.

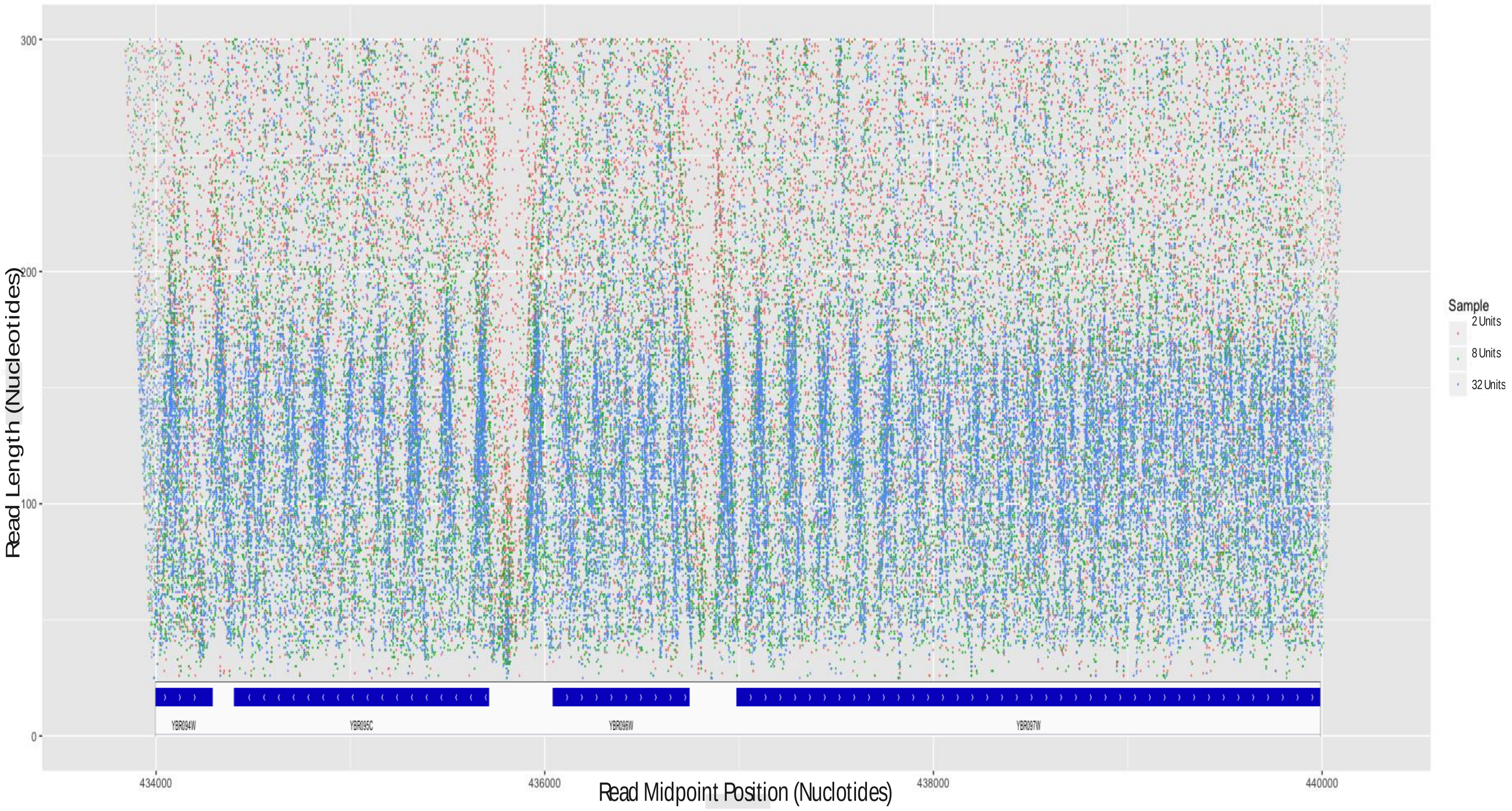


Figure 28. Midpoint position of reads in the bleomycin dataset graphed as nucleotide position vs read length combined. This is located on chromosome 2 between nucleotides 434,000 and 440,000. This is overlaid with a track showing gene position, name and orientation below it.

Because each digestion level is unique in its information it is valuable to merge the data respective to its reagent onto one graph (Figures 27 and 28). These have been aligned to an IGV gene track and we see that the data footprints have a biological basis in specified genomic regions. For example, the transcription factor footprint at 436000 is just upstream of the YBR095C which would be a TSS region. These TSS regions are surrounded by two well positioned nucleosomes on each side which is previously described in the literature. It can be seen in both figures the farther a nucleosome is from a boundary, in this case a TSS, that the less well positioned it is eventually resulting in no resolution at all in the graphs as it approaches the 440000. This is a phenomenon referred to as a stochastic model for nucleosome sliding.[19] Since there is little biological consequence for nucleosome position in the described region, we see that the positions are not consistent across the cell culture sampled and the result is a cluster at nucleosome size fragments that cannot be resolved to single nucleosomes. An interesting contrast between the two libraries is the different characterization of the middle three nucleosomes on the YBR096W gene. In the MNase library they appear to be "fuzzy" nucleosomes, however in the Bleomycin datasets they appear as resolved and single positioned nucleosomes. It is important to distinguish that "fuzzy" nucleosomes are nucleosomes with multiple positions in a region and not simply poor resolution in data. This contrast is likely reflective of the fact that MNase has difficulty digesting closely positioned nucleosomes.

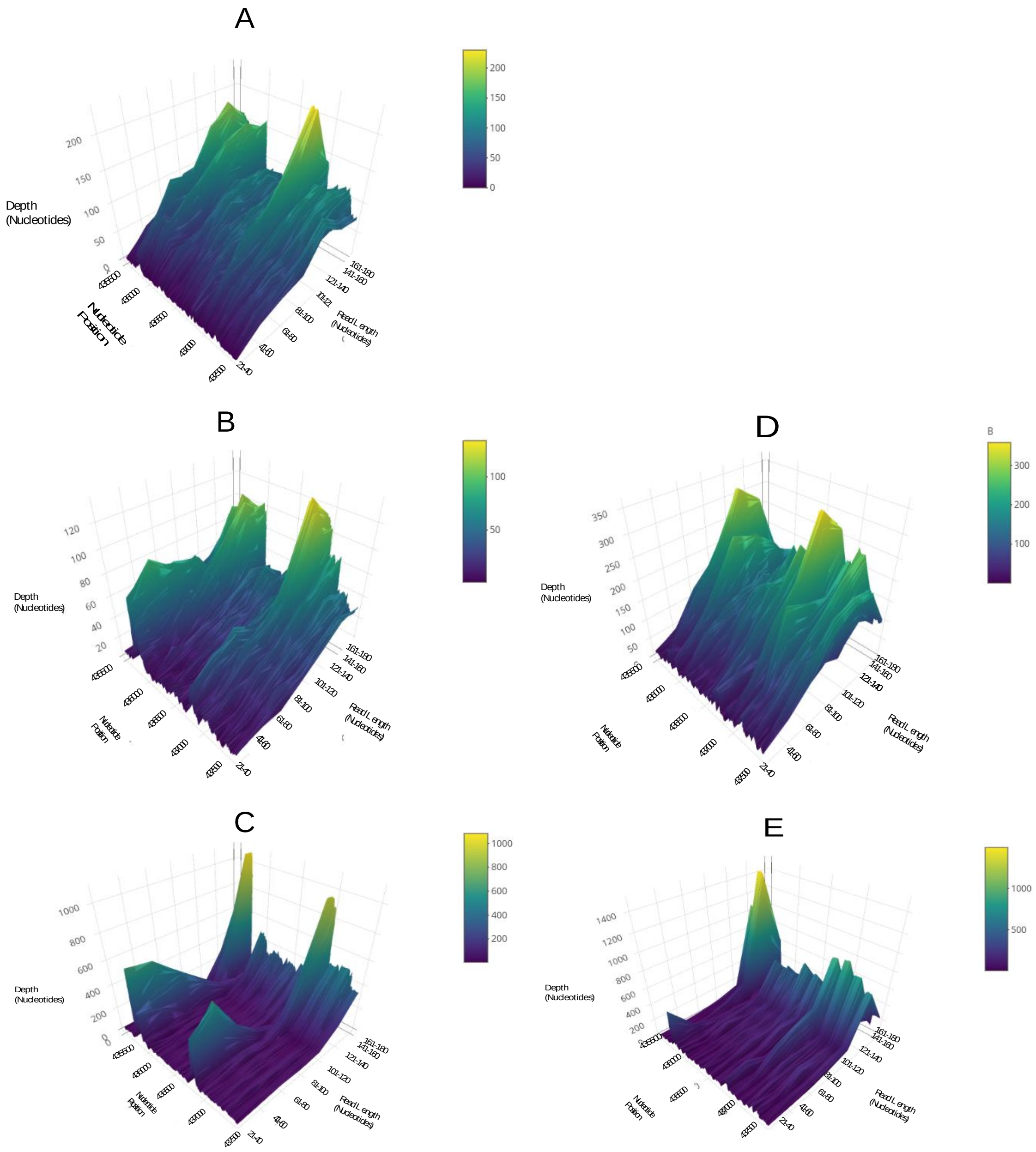


Figure 29. Surface plot representation of bleomycin data. The samples were separated by read length along the z-axis by using IGV-count to get a 1 bp sliding window of coverage at each position. These were the class sizes were then aggregated into a matrix and a surface plot was created using plot.ly. This region is 345500 to and 347500 nucleotides on the second chromosome. This represents YBR096W and the promoters barricading it. The corresponding sample order is as follows: A. 8 Units Bleomycin, B. 2 Units Bleomycin, C. 32 Units Bleomycin, D. 2.5 Units MNase, E. 20 Units MNase.

While a V-plot is a great tool to analyze chromatin data it fails to create a 3-D representation that a surface plot can **(Figure 29).** In this figure, looking at C and E, MNase has a much stronger signal to noise representation of the data that represents transcription factors. As shown in **Figure 27**, the vast majority of the data that represents nucleosomes is contained in the 161-180 nucleotide size class. This also shows that even in high digestion, shown in **Figure 29E**., MNase is not as efficient as Bleomycin as digesting linker DNA in tightly packed regions. This in shown from 436000 to 436500 as the peaks never drop to zero and are visualized as one ridge. In the Bleomycin data set however these same nucleosomes are resolved at individual positions. Comparing the Bleomycin and MNase data in these figures the two provide similar information on the chromatin structure. However, there is a stark difference in the profile and nature of the data itself. Looking at the data along the Z-axis we can see that the nucleosome data in Bleomycin spans from roughly 150 nucleotides down to 50 nucleotides. The variation here extends much lower than the MNase datasets and begins to suggest that Bleomycin can digest DNA on the nucleosome as these regions are still high resolution. In respect the concentration of depth at higher nucleotide positioning in MNase datasets could be a result of the presence of linker DNA as well.

An Assessment of Sequence Bias

A key part of this investigation is to determine if Bleomycin has a nucleotide bias. To resolve this a python script was executed to loop through the first 25 base pairs of every read in the data sets keeping a count of each nucleotide at each position. A histogram was made of this data MNase has a bias for digesting at adenine and thymine nucleotide and this is represented in **Figure 30** by the massive disproportion at position one in the MNase data. This same bias is not found when looking that first position in Bleomycin. This is reflected in **Figures 31 and 32** as there is inconsistency in codon representation across MNase datasets in that the molecule is only cutting at codon that start with adenine or thymine. While Bleomycin has a non-random selection for triplet sequences the distribution in **Figure 32** is much more dispersed than the distribution in **Figure 31**. This analysis only works as a surface level comparison however as codon bias is common phenomenon in organisms would require a normalization to the representation the entire genome. [49] However, at a base level comparison the two bar graphs have staunch differences in distribution of nucleotide triplets.

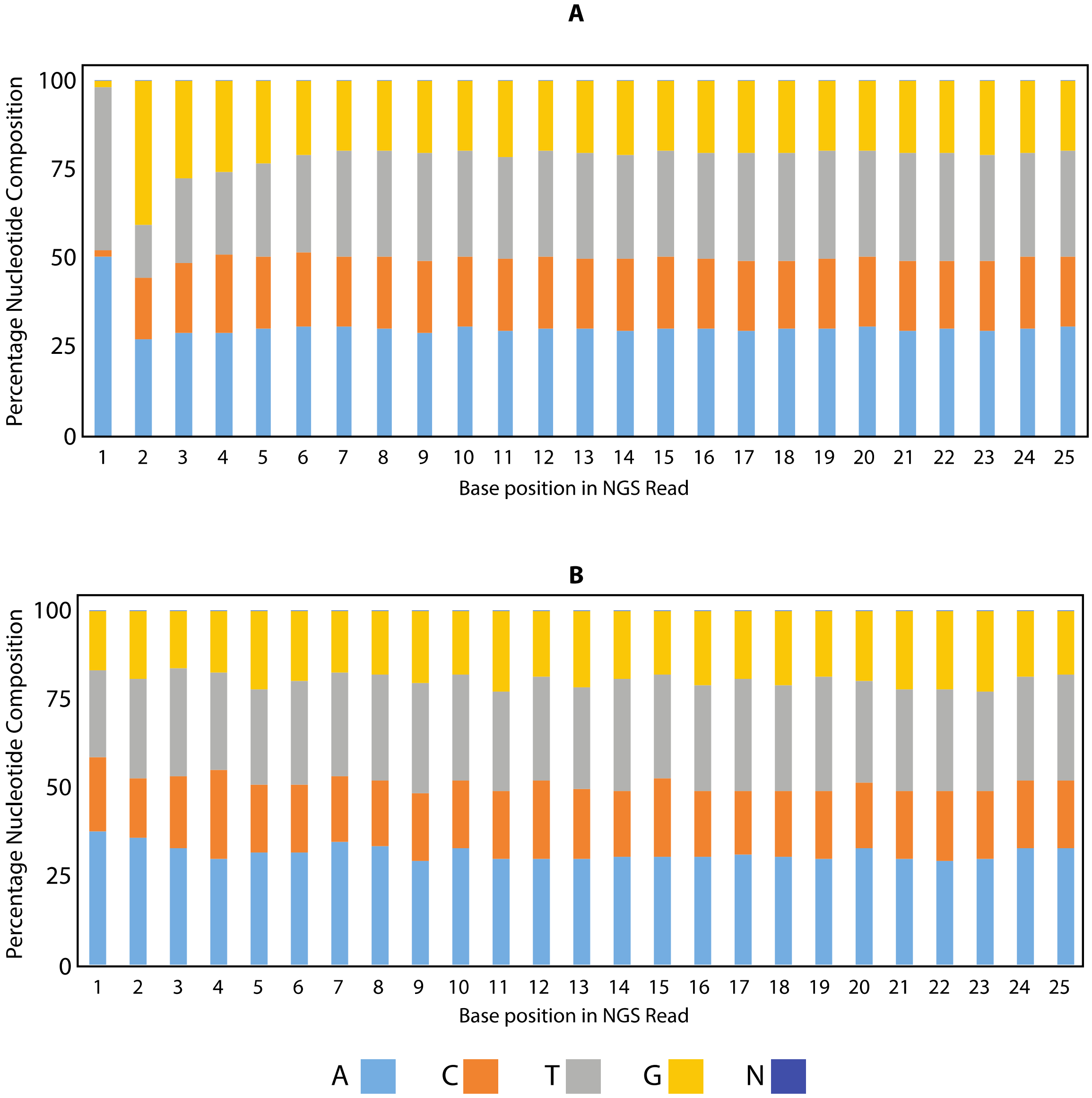


Figure 30. Stacked bar plots showing proportion of nucleotides that are at each position in the first 25 bases of each dataset running from 5' to 3'. A. 20 Units MNase Library B. 32 Units Mnase Library.

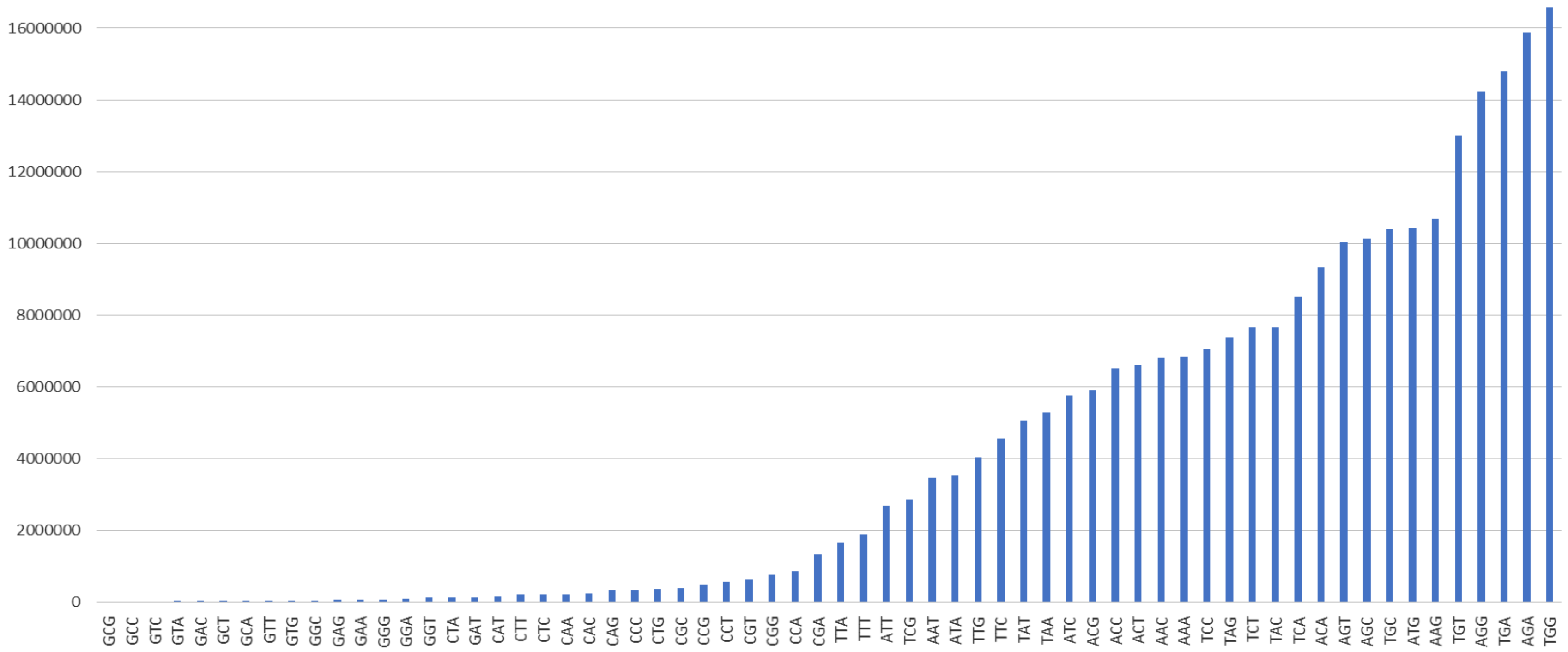


Figure 31. Bar graph of triplet sequences at the start position of reads in MNase 20 Units digestion. The number of times each triplet occurred was counted and plotted in a bar chart. The x-axis representing each triplet and the y-axis its corresponding count.

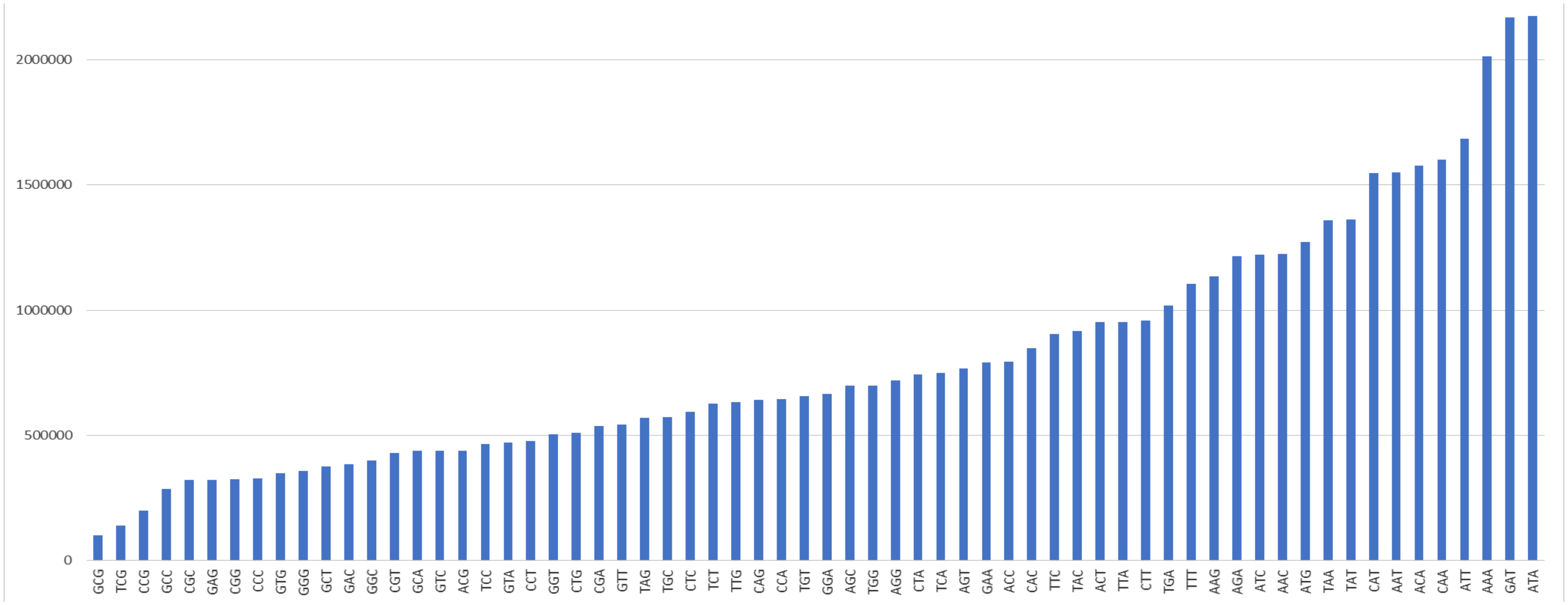


Figure 32. Bar graph of triplet sequences at the start position of reads in bleomycin 32 Units digestion. The number of times each triplet occurred was counted and plotted in a bar chart. The x-axis representing each codon and the y-axis its corresponding count.

Correlation Between Bleomycin and MNase Libraries

The correlation matrix shows that the datasets can be clustered by class size given the chromatin information present in each class **(Figure 33)**. This coincides with that transcription factors and nucleosomes are in different size classes in the datasets **(Figures 17 & 18)**. The datasets also mostly separate by reagent used as most of the MNase data clusters together at the bottom corner of the matrix. MNase does not sort as well by size in the middle size classes. This could be because the intermediate class sizes do not have many reads and as a result are not informative **(Figure 17)**. This confirms what is revealed in MNase surface plot that intermediate-sized classes have very little depth at all in MNase digestion **(Figure 29).** Bleomycin data sets however seem to striate by class size even in middle data sets implying Bleomycin can produce reads that represent nucleosomes smaller than 147. Bleomycin datasets also cluster by digestion level well implying that each level of digestion obtains unique data profiles. When creating experiments, it’s best to produce distinct and informative data and the entire lower right corner of the matrix is neither. This is concerning in that it indicates that most of the MNase data is either uninformative or not unique in signature.

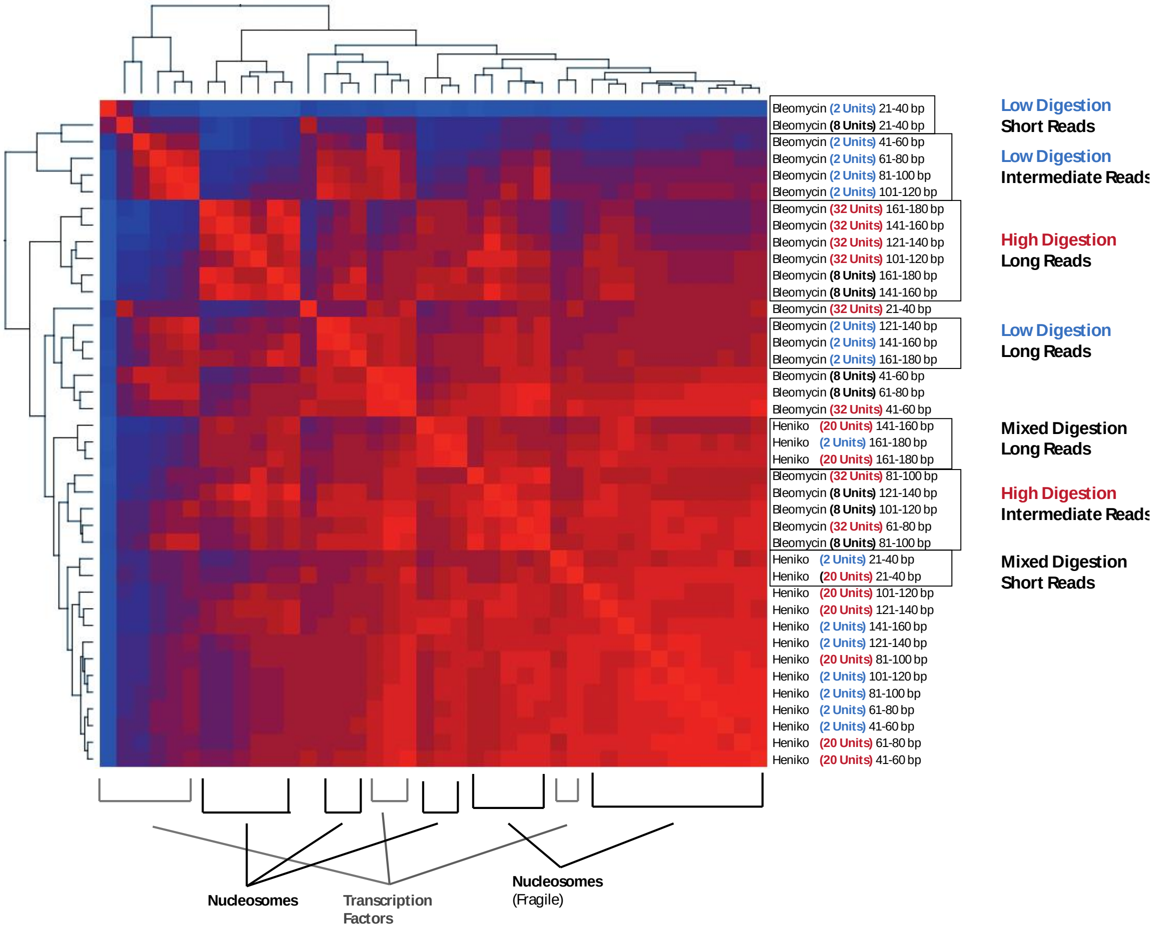


Figure 33. Correlation matrix of class sized data sets of both the Bleomycin and MNase libraries. This data was analyzed by 20 nucleotide step to create a list of coverage and position that were plotted against each other in regression analysis. Using this to create a distance matrix the samples were hierarchical clustered.

Chapter 4

Discussion

Bleomycin Can Effectively Perform Chromatin Digestion

The most fundamental question of this pursuit was, "Can Bleomycin perform chromatin digestion?" There is substantial evidence in this thesis that shows patterns of nucleosome digestion and even transcription factor representation. On a base level, gel electrophoresis illustrates that the banding pattern present after digestion reflects the weight of mono-nucleosomes, di-nucleosomes, and tri-nucleosomes **(Figure 14)**. This confirms what was stated by Kuo and Hsu that Bleomycin was one of the few non-enzymatic molecules that can be used to study chromatin structure.[41] Additionally evident in the electrophoresis patterns, as the digestion level increased, we saw a larger concentration in the band that would represents mono-nucleosomes. This follows previous biological predications as in the Henikoff study there were fewer tri-nucleosomes and di-nucleosomes in higher digestion. This can be explained by increased exposure to the reagent resulting in more linker DNA being broken down.

The read fragments that were present in the Bleomycin data set were shown to be informative as to nucleosome position with the Danpos2 analysis. If the fragments had been no digestion pattern it could be expected that there would be a straight line across all graphs in **Figures 16 and 17**. In contrast, in **Figure 18** when the reads were aligned to a TSS, we see a consistent wave pattern that when compared with literature can be understood to be signatures of transcription factors followed by well-positioned nucleosomes. This is further confirmed by the 141-160 size class being the most enriched in nucleosomes. We also evidenced the emergence of patterned waves in **Figure 16** when the reads we aggregated and centered on a well-positioned nucleosome. These "waves" represent a pattern of a nucleosome followed by linker DNA. The stochastic model of nucleosome positioning is represented in these figures as the nucleosomes are less well positioning moving away from a well-positioned nucleosome consecutively. [19]

Additionally, we can see that DNA fragment length decreases around 147 base pairs suggesting that Bleomycin is digesting at linker DNA **(Figure 18)**. These patterns all go to show that Bleomycin is digesting at linker-DNA and leaving chromatin structures in the dataset. In the case of 2 Units of Bleomycin we see a peak at 25-30 base pairs representing transcription factors. It was expected that low digestion would give the best results for transcription factor presence as this was the case for the MNase dataset. A ten base pair periodicity is present can be seen in the data and becomes more pronounced in higher digestion levels. This periodicity dissipates past 147 base pairs indicating that it is a phenomenon that does not occur in linker DNA. While further experimentation would need to be done, we can hypothesize that this periodicity corresponds to DNA having a ten base pair turn and Bleomycin preferring DNA in a specified position on the nucleosome. Bleomycin has been shown to be stabilized in the surrounding backbone of the minor groove. The limited availability of the minor groove on nucleosomes could be responsible for the periodicity seen in fragment lengths. It also shows that what has previously been referred to as "protected DNA" is not protected from digestion when the reagent is Bleomycin meaning that Bleomycin is digesting DNA on the nucleosome.[35] The evidence illustrated in this thesis strongly suggests that Bleomycin can successfully be used as a reagent in chromatin digestion.

Comparison of Bleomycin vs. MNase DNA Digestion

Bleomycin also serves as an alternative to MNase digestion as a means of chromatin digestion. In comparison Bleomycin and MNase performed similarly in the Danpos2 analysis **(Figures 16&17)**, except for MNase having nucleosome positions most enriched in the 161-180 size class. Bleomycin has nucleosome positions most enriched in the 141-160 size class. The similarity between the two does however show that Bleomycin serves as a comparable reagent in function and performance. This size difference is likely due to MNase leaving some linker-DNA undigested. One of the major differences between MNase is Bleomycin is the nucleotide bias shown in **Figure 30**. Bleomycin has little to no nucleotide bias present and can be one explanation for less linker-DNA present in the dataset. James Allan published that the nucleotide bias in MNase does not create a bias in nucleosome positioning, however we found that the resulting additional linker-DNA decreases the resolution of positioning.[35] We can see that the larger reads in the MNase dataset result in merged clusters (**Figure 26**). These merged clusters appear as fragile or “fuzzy” nucleosomes when in the Bleomycin data set they are well positioned in many cases. The added digestion of nucleosomes in tightly packed chromatin leads to a better resolution in the data. In contrast while the transcription factors are noticeable in the Bleomycin dataset they appear over digested. When this data is represented in a 3-D surface plot we can see that the two reagents are interacting with DNA in very different. At the position of nucleosomes, we see that the read count in MNase datasets is concentrated into the highest size classes and has some variation in positioning. The peaks for Bleomycin are mostly from 50-150 bp and are relatively have less variation in positioning resulting in narrower and more distinguished peaks. While both datasets are different in signature, both can clearly resolve most chromatin structures in the region selected.

Development of Novel Methods for Visualization of Chromatin Patterns

The complexity of chromatin and next-generation sequencing data leaves the task of having to generate visualization that can effectively communicate the nuance of the results. Typically, when reading these studies an area of interest is visualized using multiple coverage tracks out of IGV. One unforeseen benefit to IGV is that when the

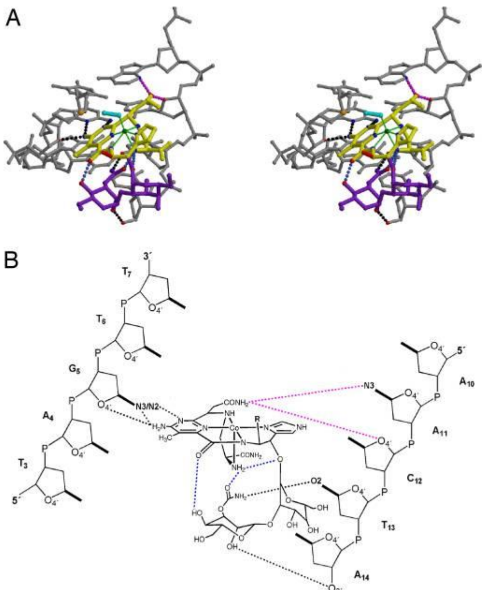


Figure 34. Interactions of the metal-binding domain and disaccharide with the DNA. (*A*) Diagram of DNA binding interacting with the metal binding and disaccharide regions of bleomycin. (*B*)Schematic showing hydrogen bonds between bleomycin and the minor groove of DNA. [50]

coverage-track views a paired-end read it only counts the ends giving insight into the location of digestion of the reagent. This can quantify nucleosome position it requires multiple figures for a single dataset. The use of a V-plot for entire regions of the genome allows to have all size classes in a single figure. In **Figures 28 and 27** the data is clear enough that it is visually easy to see the clusters and trends in the data. However, even in areas where over plotting is an issue, a 2-D density plot can be made to display clusters **(Figure 21)**. Furthermore, we even see clear visualization of a fragile nucleosome by the smearing of the first nucleosome position. Even this however does not capture the 3-Dimensional nature of chromatin as it exists. We created a 3-D dimensional represent of depth at positions in the genome at all size classes. The benefits of representing the data the way it is in **Figure 29** is that all the data are on one graph, but it is still size classed along the z-axis. Additionally, there is the emergence of three-dimensional edges to nucleosomes that give a more informative understanding to their shape and position.

The results of this experiment have raised questions, as that will need to be explored in further experiments. A molecular dynamics study was done that showed DNA stabilizes the binding of Bleomycin in the minor groove through hydrogen bonding with the backbone shown in **Figure 34**.[50] This studies limitation has in application to our work because it is done on naked DNA. [34] A repetition of this experiment could be done with the crystal structure of a nucleosome to determine if the limited availability of the minor groove gives bleomycin a preference to certain positions on the nucleosome. The results of this experiment could be cross-examined with the ten base-pair periodicities in the fragment length histogram in **Figure 14**.

One clear drawback to the Bleomycin being an aggressive DNA digestion reagent is that it over digested the data that would correspond to transcription factor fragments. Repeating the experiment by lowering the lowest tier of digestion may give a stronger signal in for transcription factor structures. Another possible source of this reduced transcription factor footprint is that a portion of the transcription factor data could have been lost in the data cleansing process in two main steps. The first is that the molecular techniques used in this experiment maintain accuracy past 125 nucleotides, however we ligated a 100-nucleotide sequence to each fragment. This means that all reads that came back smaller than 25 base pairs were filtered out of from the libraries for accuracy. As we saw in **Figure 14** the Bleomycin tends to digest chromatin structures down past what is known as their traditional size in nucleotides. Transcription factors on average are 20-60 base pairs and as a result it is likely that a portion of the transcription factor associated fragments were filtered out in this data cleansing step. The second step was filtering for reads with the highest mapping scores. While this mostly filters out errors and it has a bias toward filtering short reads with a higher probability of mapping multiple places. The MNase data was put through the same data cleansing process but because the reads are longer than expected these filters may not have impacted the signal in the data in the same way.

Future Studies

As the field continues to advance in the information age it will be necessary to utilize deep learning tools to create nuance and accurate models from large datasets. The use of neural networks has been used in the past to predict phenotypes while including epigenetic structures. Since the encroachment of nucleosomes onto the transcriptional start site is directly correlated with gene suppression **(Figure 4)**, the datasets of Bleomycin can be used to develop neural networks using nucleosome positioning and can be trained on standard datasets of gene expression. This can be repeated for MNase as a comparative study. When the predictive power is tested on known samples it can be determined which dataset is the most informative. It can be hypothesized that because Bleomycin is able to resolve heterochromatin better it will have higher predictive power in determining phenotypes or gene expression levels.

One major barrier in chromatin studies is the necessity to cross-link DNA. This is done by covalently bonding DNA with certain and preventing DNA replication processes. One inaccuracy caused by cross-linking is it can impact chromatin interactions with reagents and antibodies. [51] The success of Bleomycin as a digestive agent offers a huge application as Bleomycin primarily functions as a cancer and has been engineered for cell permeability. This offers the opportunity to digest chromatin in vivo and eliminate the need for cross-linking the sample. When no fixation is needed cells are seen to be in a more “native” state and allow for a more accurate form of study. This would allow for the experiment to be repeated in a cancer cell line such as HeLa cells and achieve permeability. In the future a comparative study could be done on separate live tissues and compare nucleosome positioning and expression levels between separate tissues.

The available tools for chromatin digestion visualization in IGV are insufficient. The hope of providing novel ways to view and communicate chromatin digestion is to build the basis of a new, guided user interface (GUI) that would implement these strategies. The benefit of having analysis tracks that using a V-plot approach is that the data does not have to be size sorted and can all be placed on one plot. To view that same figure in the coverage track we would need to open a histogram for each individual size class. The three-dimensional plots area made from a matrix of the merged size classes but can be designed to not need size classes at all. Building a streamlined GUI would allow for these approaches to be applied to the entire genome. This would allow the field to have a standardized way of effectively communicating nucleosome position.

Batch effects occur in bioinformatics when comparing different samples processed at different times and with different machines. As a result, there could be uncontrolled errors between samples that lead to false conclusions. Here we are comparing three separate batches of data; MNase data, the preliminary libraries for Bleomycin, and the deep-sequencing libraries with higher depth for Bleomycin. To minimize batch effects the libraries were analyzed separately and compared on a very broad scale. The differences in nucleotide composition were pronounced enough to overcome concerns of batch effects between MNase and Bleomycin libraries.[52]

Chapter 5

Conclusion

A method using Bleomycin to digest chromatin at single nucleotide resolution has been established in this thesis. The pharmacokinetic and pharmacodynamic optimization of this drug makes for an ideal candidate for in vivo studies of chromatin structure. The lack of a preferential nucleotide digestion bias in Bleomycin also gives a better profile of read distribution that is more reflective of nucleosome positioning than with MNase. The higher resolution of data a tightly packed chromatin positions allowed for better visualization of nucleosome positioning and gives hope for optimistic results for the digestion of heterochromatin with Bleomycin. Our results show that Bleomycin is a viable alternative to MNase digestion of chromatin structures.

References

1. Hershey A, Chase M (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage (PDF). J Gen Physiol. 36 (1): 39-56. doi:10.1085/jgp.36.1.39. PMC 2147348. PMID 12981234.

2. Ido S., Kimura K., Oyabu N., Kobayashi K., Tsukada M., Matsushige K., and Yamada H. (2013). Beyond the Helix Pitch: Direct Visualization of Native DNA in Aqueous Solutions. ACS Nano 7 (2), 1817-1822, DOI: 10.1021/nn400071n

3. Watson J. & Crick F. (1953). Molecular Structure of Nucleic Acids. A Structure for Deoxyribose Nucleic Acid., Nature (3), 171, 737-738

4. Poveda, A. M., Le Clech, M., & Pasero, P. (2010). Transcription and replication: breaking the rules of the road causes genomic instability. Transcription, 1(2), 99-102.

5. Rose GD, Fleming PJ, Banavar JR, Maritan A (November 2006). A backbone-based theory of protein folding. Proceedings of the National Academy of Sciences of the United States of America. 103 (45): 16623–33. Bibcode:2006PNAS..10316623R. CiteSeerX 10.1.1.630.5487. doi:10.1073/pnas.0606843103. PMC 1636505. PMID 17075053.

6. Fersht A (1999). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. Macmillan. ISBN 978-0-7167-3268-6.

7. Hardison, R. C. (2012-12-01). Evolution of hemoglobin and its genes. Cold Spring Harbor Perspectives in Medicine. 2(12): a011627. doi:10.1101/cshperspect.a011627. ISSN 2157-1422. PMC 3543078. PMID 23209182.

8. Björklund, N.K. & R. Gordon. (1993) Nuclear state splitting: a working model for the mechanochemical coupling of differentiation "waves" with the controlling genes (master genes) [Russian]. Ontogenez 24(2), 5-23

9. Noble, D. (2015) Conrad Waddington and the Origin of Epigenetics. Journal of Experimental Biology, vol. 218, no. 6, pp. 816–818., doi:10.1242/jeb.120071.

10. Ng H.H., Robert F., Young R.A., Struhl K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Molecular Cell. 11 (3): 709–19. doi:10.1016/S1097-2765(03)00092-3. PMID 12667453.

11. Reece J., Campbell N. (2006). Biology. San Francisco: Benjamin Cummings. ISBN 978-0-8053-6624-2.

12. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 389 (6648): 251–60. Bibcode:1997 Nature,.389..251L. doi:10.1038/38444. PMID 9305837.

13. Sivaraman, P., & Kumarevel, T. (2015). Crystal structure of the nucleosome core particle. Nature. doi:10.2210/pdb3x1s/pdb

14. Wheeler, R. (2013). Chromatin Structures. Wikipedia

15. Khorasanizadeh S., (2004) The Nucleosome: From Genomic Organization to Genomic Regulation, Cell, Volume 116, Issue 2,Pages 259-272,ISSN 0092 8674,https://doi.org/10.1016/S0092-8674(04)00044-3.(http://www.sciencedirect.com/science/article/pii/S0092867404000443)

16. Christophorou M.A., Castelo-Branco G., Halley-Stott R.P., Oliveira C.S., Loos R., Radzisheuskaya A., Mowen K.A., Bertone P., Silva J.C., Zernicka-Goetz M., Nielsen M.L., Gurdon J.B., Kouzarides T. (2014). Citrullination regulates pluripotency and histone H1 binding to chromatin. Nature. 507 (7490): 104–8. Bibcode:2014Natur.507..104C. doi:10.1038/nature12942. PMC 4843970. PMID 24463520.

17. Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature, 458(7236), 362-6. Retrieved from http://ezproxy.rowan.edu/login?url=https://search.proquest.com/docview/204466086?accountid=13605

18. Zhang, T., Zhang, W., & Jiang, J. (2015). Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants. Plant Physiology,168(4), 1406-1416. doi:10.1104/pp.15.00125

19. Padinhateeri, R., & Marko, J. F. (2011). Nucleosome positioning in a model of active chromatin remodeling enzymes. Proceedings of the National Academy of Sciences,108(19), 7799-7803. doi:10.1073/pnas.1015206108

20. Bassel J. and Mortimer R. (1971) Genetic order of the galactose structural genes in Saccharomyces cerevisiae. J Bacteriol 108(1):179-83 PMID: 5122803

21. Copland J.A., Sheffield-Moore M., Koldzic-Zivanovic N., Gentry S., Lamprou G., Tzortzatou-Stathopoulou F., Zoumpourlis V., Urban R.J., Vlahopoulos S.A. (2009). Sex steroid receptors in skeletal differentiation and epithelial neoplasia: is tissue-specific intervention possible?. BioEssays. 31 (6): 629–41. doi:10.1002/bies.200800138. PMID 19382224.

22. Ohlsson, R., & Gondor, A. (2018). Faculty of 1000 evaluation for YY1 Is a Structural Regulator of Enhancer-Promoter Loops. F1000 - Post-publication Peer Review of the Biomedical Literature. doi:10.3410/f.732279386.793546049

23. Jabbari K, Bernardi G (2004). Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 333: 143–9. doi:10.1016/j.gene. PMID 15177689.

24. Fatemi M., Pao M.M., Jeong S., Gal-Yam E.N., Egger G., Weisenberger D.J., Jones P.A. (2005). "Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level". Nucleic Acids Res. 33 (20): e176. doi:10.1093/nar/gni180. PMC 1292996. PMID 16314307.

25. Illingworth R.S., Gruenewald-Schneider U., Webb S., Kerr A.R., James K.D., Turner D.J., Smith C., Harrison D.J., Andrews R., Bird A.P. (2010). "Orphan CpG islands identify numerous conserved promoters in the mammalian genome". PLoS Genet. 6 (9): e1001134. doi:10.1371/journal.pgen.1001134. PMC 2944787. PMID 20885785.

26. Mandal S.S. (2010). "Mixed lineage leukemia: versatile player in epigenetics and human disease". The FEBS Journal. 277(8): 1789. doi:10.1111/j.1742-4658.2010.07605.x. PMID 20236314.

27. Pokholok D., Harbison C., Levine S., Cole M., Hannett N., Lee T., Bell G., Walker K., Rolfe P., Herbolsheimer E., Zeitlinger J., Lewitter F., Gifford D., Young R., Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast, Cell, Volume 122, Issue 4,2005,517-527,ISS,0092-8674,https://doi.org/10.1016/j.cell.2005.06.026., (http://www.sciencedirect.com/science/article/pii/S0092867405006458)

28. Miller, N. A., Farrow, E. G., Gibson, M., Willig, L. K., Twist, G., Yoo, B., . . . Kingsmore, S. F. (2015). A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Medicine,7(1). doi:10.1186/s13073-015-0221-8

29. National Human Genome Research Institute (NHGRI) Homepage | NHGRI. (n.d.). Retrieved from https://www.genome.gov/

30. Mardis, E. R. (2006). "Anticipating the 1,000 dollar genome". Genome Biology. 7 (7): 112. doi:10.1186/gb-2006-7-7-112. PMC 1779559. PMID 17224040.

31. Unlocking the Power of the Genome. (n.d.). Retrieved from https://www.illumina.com/

32. Embl-Ebi. (n.d.). The European Bioinformatics Institute. Retrieved from https://www.ebi.ac.uk/

33. An integrated encyclopedia of DNA elements in the human genome. (2012). Nature, 489(7414), 57-74. Retrieved from http://ezproxy.rowan.edu/login?url=https://search.proquest.com/docview/1069238769?accountid=13605

34. Chung H.R., Dunkel I., Heise F., Linke C., Krobitsch S., Article Source: The Effect of Micrococcal Nuclease Digestion on Nucleosome Positioning Data (2010) PLOS ONE 5(12): e15754. https://doi.org/10.1371/journal.pone.0015754

35. Allan, J., Fraser, R. M., Owen-Hughes, T., & Keszenman-Pereyra, D. (2012). Micrococcal Nuclease Does Not Substantially Bias Nucleosome Mapping. Journal of Molecular Biology,417(3), 152-164. doi:10.1016/j.jmb.2012.01.043

36. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M., & Henikoff, S. (2011). Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(45), 18318–18323. doi:10.1073/pnas.1110731108

37. https://zh.wikibooks.org/wiki/File:Bleomycin_SAR.svg

38. Einhorn, L.H. Curing metastatic testicular cancer. PNAS 99, 4592-4595 (2002).

39. Sigma Aldrich. Bleomycin Sulfate from Streptomyces verticillus. in Product Information Vol. 2016 (Sigma Aldrich).

40. Hecht, Sidney M., Bleomycin: New Perspectives on the Mechanism of Action, 2000/01/01, doi: 10.1021/np990549f, Journal of Natural Products, https://doi.org/10.1021/np990549f

41. Kuo, M.T. & Hsu, T.C. Bleomycin causes release of nucleosomes from chromatin and chromosomes. Nature 271, 83-84 (1978).

42. Langmead, Ben; Cole Trapnell; Mihai Pop; Steven L Salzberg (4 March 2009). *"Ultrafast and memory-efficient alignment of short DNA sequences to the human genome"* (PDF). *Genome Biology*. 10 (3): 10:R25. doi:10.1186/gb-2009-10-3-r25. PMC 2690996. PMID 19261174. Retrieved 29 November 2013.

43. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools" (PDF). *Bioinformatics*. **25** (16): 2078–2079. doi:10.1093/bioinformatics/btp352. PMC 2723002. PMID 19505943.

44. Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov.
Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**.** Briefings in Bioinformatics 14, 178-192 (2013).
45. Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome research*, *23*(2), 341–351. doi:10.1101/gr.142067.112

46. Shen, L., Shao, N., Liu, X., & Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, *15*(1), 284. doi: 10.1186/1471-2164-15-284

47. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

48. Li G., Levitus M., Bustamante C., Widom J. Rapid spontaneous accessibility of nucleosomal DNA. Nat Struct Mol Biol. 2005;12:46–53.

49. Athey J., Alexaki A., Osipova E., Rostovtsev A., Santana-Quintero L., Katneni U., Simonyan V., Kimchi-Sarfaty C. (2017-09-02). "A new and updated resource for codon usage tables". *BMC Bioinformatics*. **18** (391): 391. doi:10.1186/s12859-017-1793-7. PMC 5581930. PMID 28865429.

50. Goodwin, K. D., Lewis, M. A., Long, E. C., & Georgiadis, M. M. (2008). Crystal structure of DNA-bound Co(III) bleomycin B2: Insights on intercalation and minor groove binding. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5052–5056. doi:10.1073/pnas.0708143105

51. Chromatrap, and Porvair Sciences Ltd. (2016). Advantages and Disadvantages of Native and Cross-Linked Chromatin Immunoprecipitation. *News*, 16, www.news-medical.net/whitepaper/20160616/Advantages-and-Disadvantages-of-Native-and-Cross-Linked-Chromatin-Immunoprecipitation.aspx.

52. Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrens, T. W., & Bhangale, T. R. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics*, *18*(1). doi: 10.1186/s12859-017-1756-z

Appendix A: Sample Pipeline For Formatting and Cleansing

```
# set -e makes program stop if line of code failed
set -e
# Pulling out Unique Reads

head -n 19 GSM756842.sam > GSM756842.header

grep YT:Z:CP GSM756842.sam > GSM756842.clean.sam

#cat Mnase_Sperm_Unspun_NoIndex_L004_R1_002.fastq >>
Mnase_Sperm_Unspun_R1.fastq


#Making Bam file format

cat GSM756842.clean.sam >> GSM756842.header
rm GSM756842.clean.sam
mv GSM756842.header GSM756842.clean.sam

#Making Bam file
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o
GSM756842.clean.bam GSM756842.clean.sam

#Sorting Bam File


#/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o
GSM756842.clean.sorted.bam GSM756842.clean.bam


# Perl scripts to see the size distribution of library

#perl -e ' $col=8;  while (<>) { s/\r?\n//; @F = split /\t/, $_; $val = $F[$col]; if (! exists
$count{$val}) { push @order, $val } $count{$val}++; } foreach $val (@order) { print
"$val\t$count{$val}\n" } warn "\nPrinted number of occurrences for ", scalar(@order), "
values in $. lines.\n\n"; ' Mnase_Sperm_Unspun_unique.sam >
Mnase_Sperm_Unspun_unique.counts.txt



#head -n 24 Mnase_Sperm_Unspun_unique.sam > Mnase_Sperm_Unspun.header



#Making Bam files for 135-165
```

```
#awk ' $9 <= 165 && $9 >= 135 || $9 >= -165 && $9 <= -135 '
Mnase_Sperm_Unspun_unique.sam > Mnase_Sperm_Unspun.135-165.sam
#cp Mnase_Sperm_Unspun.header Mnase_Sperm_Unspun.135-165.header
#cat Mnase_Sperm_Unspun.135-165.sam >> Mnase_Sperm_Unspun.135-165.header
#rm Mnase_Sperm_Unspun.135-165.sam
#mv Mnase_Sperm_Unspun.135-165.header Mnase_Sperm_Unspun.135-165.sam
#/share/bin/samtools/samtools view -S -t /home/caroneb/scratch/bigbed/mm9.chrom.sizes
-b -o Mnase_Sperm_Unspun.135-165.bam Mnase_Sperm_Unspun.135-165.sam
#rm Mnase_Sperm_Unspun.135-165.sam

#Making Bam files for 100-130


#awk ' $9 <= 130 && $9 >= 100 || $9 >= -130 && $9 <= -100 '
Mnase_Sperm_Unspun_unique.sam > Mnase_Sperm_Unspun.100-130.sam
#cp Mnase_Sperm_Unspun.header Mnase_Sperm_Unspun.100-130.header
#cat Mnase_Sperm_Unspun.100-130.sam >> Mnase_Sperm_Unspun.100-130.header
#rm Mnase_Sperm_Unspun.100-130.sam
#mv Mnase_Sperm_Unspun.100-130.header Mnase_Sperm_Unspun.100-130.sam
#/share/bin/samtools/samtools view -S -t /home/caroneb/scratch/bigbed/mm9.chrom.sizes
-b -o Mnase_Sperm_Unspun.100-130.bam Mnase_Sperm_Unspun.100-130.sam
#rm Mnase_Sperm_Unspun.100-130.sam



#Making Bam files for 1-80

#awk ' $9 <= 80 && $9 >= 1 || $9 >= -80 && $9 <= -1 '
Mnase_Sperm_Unspun_unique.sam > Mnase_Sperm_Unspun.1-80.sam
#cp Mnase_Sperm_Unspun.header Mnase_Sperm_Unspun.1-80.header
#cat Mnase_Sperm_Unspun.1-80.sam >> Mnase_Sperm_Unspun.1-80.header
#rm Mnase_Sperm_Unspun.1-80.sam
#mv Mnase_Sperm_Unspun.1-80.header Mnase_Sperm_Unspun.1-80.sam
#/share/bin/samtools/samtools view -S -t /home/caroneb/scratch/bigbed/mm9.chrom.sizes
-b -o Mnase_Sperm_Unspun.1-80.bam Mnase_Sperm_Unspun.1-80.sam
#rm Mnase_Sperm_Unspun.1-80.sam

# Compress files

#gzip Mnase_Sperm_Unspun.clean.sam
#gzip Mnase_Sperm_Unspun_unique.sam
#gzip Mnase_Sperm_Unspun_R1.fastq
#gzip Mnase_Sperm_Unspun_R2.fastq
```

```
#Sorts Bam files

#/share/bin/samtools/samtools sort GFP-K4me3.bam GFP-K4me3_sorted

#Creates Bedgraph from Bam files

#genomeCoverageBed -bg -ibam GFP-K4me3_sorted.bam -g
/home/caroneb/nearline/bigbed/mm9.chrom.sizes > GFP-K4me3.bedgraph


#Creates BigWig from Bedgraph

#/home/caroneb/nearline/bigbed/bedGraphToBigWig GFP-K4me3.bedgraph
/home/caroneb/nearline/bigbed/mm9.chrom.sizes GFP-K4me3.bw

#Copy the .bw (bigwig file) to the genomedata/webroot folder on the R drive and then go
to UCSC and paste the following into submit tracks
#track type=bigWig name="GFP-K4me3_unique" description="GFP-K4me3_unique"
itemRgb="On" bigDataUrl=http://labs.umassmed.edu/randolab/genomedata/GFP-
K4me3_unique.bw



# Perl scripts to see the size distribution of library

#perl -e ' $col=8;  while (<>) { s/\r?\n//; @F = split /\t/, $_; $val = $F[$col]; if (! exists
$count{$val}) { push @order, $val } $count{$val}++; } foreach $val (@order) { print
"$val\t$count{$val}\n" } warn "\nPrinted number of occurrences for ", scalar(@order), "
values in $. lines.\n\n"; ' ES_Chip_seq_unique.sam > ES_Chip_seq_unique.counts.txt



#head -n 24 ES_Chip_seq_unique.sam > ES_Chip_seq.header
```

Appendix B: Mapping Pipeline

```
#/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2

/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_1_1_ATCACG.fastq -2 FGC1748_s_1_2_ATCACG.fastq -S
FGC1748_s_1_ATCACG.sam --dovetail --trim3 50
/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_1_1_CGATGT.fastq -2 FGC1748_s_1_2_CGATGT.fastq -S
FGC1748_s_1_CGATGT.sam --dovetail --trim3 50
/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_1_1_TTAGGC.fastq -2 FGC1748_s_1_2_TTAGGC.fastq -S
FGC1748_s_1_TTAGGC.sam --dovetail --trim3 50
/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_2_1_ATCACG.fastq -2 FGC1748_s_2_2_ATCACG.fastq -S
FGC1748_s_2_ATCACG.sam --dovetail --trim3 50
/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_2_1_CGATGT.fastq -2 FGC1748_s_2_2_CGATGT.fastq -S
FGC1748_s_2_CGATGT.sam --dovetail --trim3 50
/Users/stolzj92/Documents/Programs/bowtie2-2.3.4-macos-x86_64/bowtie2 -x
/Users/stolzj92/Desktop/Bleomycin_Cerevisiae_10_30_2017/Genomes/sacCer2/sacCer2
-1 FGC1748_s_2_1_TTAGGC.fastq -2 FGC1748_s_2_2_TTAGGC.fastq -S
FGC1748_s_2_TTAGGC.sam --dovetail --trim3 50
```

Appendix C: A Sam to Bam Pipeline

Making Bam files for 81-100

```
awk ' $9 <= 100 && $9 >= 81 || $9 >= -100 && $9 <= -81 ' TTAGGC.clean.sam>
TTAGGC.81-100.sam
head -n 19 TTAGGC.clean.sam > TTAGGC.81-100.header
cat TTAGGC.81-100.sam >> TTAGGC.81-100.header
wc TTAGGC.81-100.sam
rm TTAGGC.81-100.sam
mv TTAGGC.81-100.header TTAGGC.81-100.sam

/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o TTAGGC.81-
100.bam TTAGGC.81-100.sam
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o
TTAGGC.clean.sorted.81-100.bam TTAGGC.81-100.bam



#Making Bam files for 101-120


awk ' $9 <= 120 && $9 >= 101 || $9 >= -120 && $9 <= -101 ' TTAGGC.clean.sam>
TTAGGC.101-120.sam
head -n 19 TTAGGC.clean.sam > TTAGGC.101-120.header
cat TTAGGC.101-120.sam >> TTAGGC.101-120.header
wc TTAGGC.101-120.sam
rm TTAGGC.101-120.sam
mv TTAGGC.101-120.header TTAGGC.101-120.sam

/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o TTAGGC.101-
120.bam TTAGGC.101-120.sam
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o
TTAGGC.clean.sorted.101-120.bam TTAGGC.101-120.bam



#Making Bam files for 121-140
```

```
awk ' $9 <= 140 && $9 >= 121 || $9 >= -140 && $9 <= -121 ' TTAGGC.clean.sam>
TTAGGC.121-140.sam
head -n 19 TTAGGC.clean.sam > TTAGGC.121-140.header
cat TTAGGC.121-140.sam >> TTAGGC.121-140.header
wc TTAGGC.121-140.sam
rm TTAGGC.121-140.sam
mv TTAGGC.121-140.header TTAGGC.121-140.sam

/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o TTAGGC.121-
140.bam TTAGGC.121-140.sam
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o
TTAGGC.clean.sorted.121-140.bam TTAGGC.121-140.bam



#Making Bam files for 141-160


awk ' $9 <= 160 && $9 >= 141 || $9 >= -160 && $9 <= -141 ' TTAGGC.clean.sam>
TTAGGC.141-160.sam
head -n 19 TTAGGC.clean.sam > TTAGGC.141-160.header
cat TTAGGC.141-160.sam >> TTAGGC.141-160.header
wc TTAGGC.141-160.sam
rm TTAGGC.141-160.sam
mv TTAGGC.141-160.header TTAGGC.141-160.sam

/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o TTAGGC.141-
160.bam TTAGGC.141-160.sam
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o
TTAGGC.clean.sorted.141-160.bam TTAGGC.141-160.bam



#Making Bam files for 161-180


awk ' $9 <= 180 && $9 >= 161 || $9 >= -180 && $9 <= -161 ' TTAGGC.clean.sam>
TTAGGC.161-180.sam
head -n 19 TTAGGC.clean.sam > TTAGGC.161-180.header
cat TTAGGC.161-180.sam >> TTAGGC.161-180.header
wc TTAGGC.161-180.sam
rm TTAGGC.161-180.sam
mv TTAGGC.161-180.header TTAGGC.161-180.sam
```

```
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools view -S -o TTAGGC.161-180.bam TTAGGC.161-180.sam
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools sort -o TTAGGC.clean.sorted.161-180.bam TTAGGC.161-180.bam

#Index all of the .sorted.bam files
/Volumes/Carone_Lab_/Programs/samtools/bin/samtools index *.sorted.bam
```

Appendix D: Codon Histogram Script Sam File

```
import os
import re
import sys

nucleotide_count1 = {}
sam_file = open(sys.argv[1])
for line in sam_file:
        if line.startswith("D"):
                split = line.split("\t")
                sequence = split[9]
                codon = sequence[0:3]
                current_count = nucleotide_count1.get(codon,0)
                new_count = current_count + 1
                nucleotide_count1[codon] = new_count


for codon, count in nucleotide_count1.items():

print(codon + " : " + str(count) + ",")
```

Appendix E: Codon Histogram Script Fastq File

```
import re
import sys
from Bio import SeqIO

nucleotide_count1 = {}

fastq_file = sys.argv[1]

for record in SeqIO.parse(fastq_file, "fastq"):
    sequence = record.seq
    codon = sequence[0:3]
    current_count = nucleotide_count1.get(codon,0)
    new_count = current_count + 1
    nucleotide_count1[codon] = new_count


for codon, count in nucleotide_count1.items():
    print(codon + " : " + str(count) + ",")
```

Appendix F: Nucleotide Count Python Script

```
import os
import sys
import re
sam_file = open(sys.argv[1])
output_file = open("sam_histogram2.csv", "w")

import re
import sys
from Bio import SeqIO

nucleotide_count1 = {}
nucleotide_count2 = {}
nucleotide_count3 = {}
nucleotide_count4 = {}
nucleotide_count5 = {}
nucleotide_count6 = {}
nucleotide_count7 = {}
nucleotide_count8 = {}
nucleotide_count9 = {}
nucleotide_count10 = {}
nucleotide_count11 = {}
nucleotide_count12 = {}
nucleotide_count13 = {}
nucleotide_count14 = {}
nucleotide_count15 = {}
nucleotide_count16 = {}
nucleotide_count17 = {}
nucleotide_count18 = {}
nucleotide_count19 = {}
nucleotide_count20 = {}
nucleotide_count21 = {}
nucleotide_count22 = {}
nucleotide_count23 = {}
nucleotide_count24 = {}
nucleotide_count25 = {}

fastq_file = sys.argv[1]

for record in SeqIO.parse(fastq_file, "fastq"):
	sequence = record.seq
	nucleotide = sequence[0:25]
	for letter in nucleotide[0]:
		current_count = nucleotide_count1.get(letter,0)
		new_count = current_count + 1
```

```
    nucleotide_count1[letter] = new_count
for letter in nucleotide[1]:
    current_count = nucleotide_count2.get(letter,0)
    new_count = current_count + 1
    nucleotide_count2[letter] = new_count
for letter in nucleotide[2]:
    current_count = nucleotide_count3.get(letter,0)
    new_count = current_count + 1
    nucleotide_count3[letter] = new_count
for letter in nucleotide[3]:
    current_count = nucleotide_count4.get(letter,0)
    new_count = current_count + 1
    nucleotide_count4[letter] = new_count
for letter in nucleotide[4]:
    current_count = nucleotide_count5.get(letter,0)
    new_count = current_count + 1
    nucleotide_count5[letter] = new_count
for letter in nucleotide[5]:
    current_count = nucleotide_count6.get(letter,0)
    new_count = current_count + 1
    nucleotide_count6[letter] = new_count
for letter in nucleotide[6]:
    current_count = nucleotide_count7.get(letter,0)
    new_count = current_count + 1
    nucleotide_count7[letter] = new_count
for letter in nucleotide[7]:
    current_count = nucleotide_count8.get(letter,0)
    new_count = current_count + 1
    nucleotide_count8[letter] = new_count
for letter in nucleotide[8]:
    current_count = nucleotide_count9.get(letter,0)
    new_count = current_count + 1
    nucleotide_count9[letter] = new_count
for letter in nucleotide[9]:
    current_count = nucleotide_count10.get(letter,0)
    new_count = current_count + 1
    nucleotide_count10[letter] = new_count
for letter in nucleotide[10]:
    current_count = nucleotide_count11.get(letter,0)
    new_count = current_count + 1
    nucleotide_count11[letter] = new_count
for letter in nucleotide[11]:
    current_count = nucleotide_count12.get(letter,0)
    new_count = current_count + 1
    nucleotide_count12[letter] = new_count
for letter in nucleotide[12]:
    current_count = nucleotide_count13.get(letter,0)
    new_count = current_count + 1
    nucleotide_count13[letter] = new_count
for letter in nucleotide[13]:
    current_count = nucleotide_count14.get(letter,0)
    new_count = current_count + 1
    nucleotide_count14[letter] = new_count
for letter in nucleotide[14]:
    current_count = nucleotide_count15.get(letter,0)
    new_count = current_count + 1
    nucleotide_count15[letter] = new_count
for letter in nucleotide[15]:
    current_count = nucleotide_count16.get(letter,0)
    new_count = current_count + 1
    nucleotide_count16[letter] = new_count
for letter in nucleotide[16]:
    current_count = nucleotide_count17.get(letter,0)
    new_count = current_count + 1
    nucleotide_count17[letter] = new_count
for letter in nucleotide[17]:
    current_count = nucleotide_count18.get(letter,0)
    new_count = current_count + 1
    nucleotide_count18[letter] = new_count
for letter in nucleotide[18]:
    current_count = nucleotide_count19.get(letter,0)
    new_count = current_count + 1
    nucleotide_count19[letter] = new_count
for letter in nucleotide[19]:
    current_count = nucleotide_count20.get(letter,0)
    new_count = current_count + 1
    nucleotide_count20[letter] = new_count
for letter in nucleotide[20]:
    current_count = nucleotide_count21.get(letter,0)
    new_count = current_count + 1
    nucleotide_count21[letter] = new_count
for letter in nucleotide[21]:
    current_count = nucleotide_count22.get(letter,0)
    new_count = current_count + 1
    nucleotide_count22[letter] = new_count
for letter in nucleotide[22]:
    current_count = nucleotide_count23.get(letter,0)
    new_count = current_count + 1
    nucleotide_count23[letter] = new_count
for letter in nucleotide[23]:
    current_count = nucleotide_count24.get(letter,0)
    new_count = current_count + 1
    nucleotide_count24[letter] = new_count
for letter in nucleotide[24]:
    current_count = nucleotide_count25.get(letter,0)
    new_count = current_count + 1
    nucleotide_count25[letter] = new_count

for letter, count in nucleotide_count1.items():
print(letter + " : " + str(count) + ",")
print(1)
for letter, count in nucleotide_count2.items():
print(letter + " : " + str(count) + ",")
print(2)
for letter, count in nucleotide_count3.items():
print(letter + " : " + str(count) + ",")
print(3)
for letter, count in nucleotide_count4.items():
print(letter + " : " + str(count) + ",")
print(4)
for letter, count in nucleotide_count5.items():
print(letter + " : " + str(count) + ",")
print(5)
for letter, count in nucleotide_count6.items():
print(letter + " : " + str(count) + ",")
print(6)
for letter, count in nucleotide_count7.items():
print(letter + " : " + str(count) + ",")
print(7)
for letter, count in nucleotide_count8.items():
print(letter + " : " + str(count) + ",")
print(8)
for letter, count in nucleotide_count9.items():
print(letter + " : " + str(count) + ",")
print(9)
for letter, count in nucleotide_count10.items():
print(letter + " : " + str(count) + ",")
print(10)
for letter, count in nucleotide_count11.items():
print(letter + " : " + str(count) + ",")
print(11)
for letter, count in nucleotide_count12.items():
print(letter + " : " + str(count) + ",")
print(12)
for letter, count in nucleotide_count13.items():
print(letter + " : " + str(count) + ",")
print(13)
for letter, count in nucleotide_count14.items():
```

```
            print(letter + " : " + str(count) + ",")
print(14)
for letter, count in nucleotide_count15.items():
            print(letter + " : " + str(count) + ",")
print(15)
for letter, count in nucleotide_count16.items():
            print(letter + " : " + str(count) + ",")
print(16)
for letter, count in nucleotide_count17.items():
            print(letter + " : " + str(count) + ",")
print(17)
for letter, count in nucleotide_count18.items():
            print(letter + " : " + str(count) + ",")
print(18)
for letter, count in nucleotide_count19.items():
            print(letter + " : " + str(count) + ",")
print(19)
for letter, count in nucleotide_count20.items():
            print(letter + " : " + str(count) + ",")
print(20)
for letter, count in nucleotide_count21.items():
            print(letter + " : " + str(count) + ",")
print(21)
for letter, count in nucleotide_count22.items():
            print(letter + " : " + str(count) + ",")
print(22)
for letter, count in nucleotide_count23.items():
            print(letter + " : " + str(count) + ",")
print(23)
for letter, count in nucleotide_count24.items():
            print(letter + " : " + str(count) + ",")
print(24)
for letter, count in nucleotide_count25.items():
            print(letter + " : " + str(count) + ",")
print(25)
```

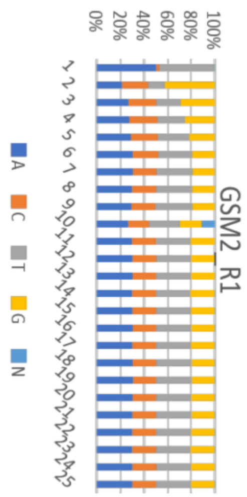
Appendix G: V-plot Script R

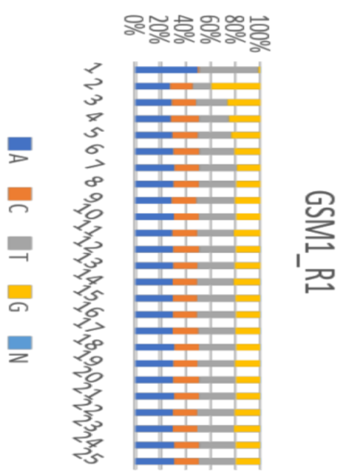
```
#input the path of the sam file at the command line
histogram = {}
for line in sam_file:
        if line.startswith("D"):
                split = line.split("\t")
                length = split[8]
                current_count = histogram.get(length, 0)
                new_count = current_count + 1
                histogram[length] = new_count
for length, count in histogram.items():
        output_file.write(length + " * " + str(count) + ",")
startpositiongsm2 <- read.csv("Desktop/startpositiongsm2.csv", sep = ",", header = F)
startpositiongsm1 <- read.csv("Desktop/startpositiongsm1.csv", header = F)
endpositiongsm2 <- read.csv("Desktop/endpositiongsm2.csv", header = F)
endpositiongsm1 <- read.csv("Desktop/endpositiongsm1.csv", header = F)
vplotGSM1 <- read.csv("Desktop/vplotGSM1.csv", header = F)
vplotGSM2<- read.csv("Desktop/vplotGSM2.csv", header = F)
ggplot(DT, aes(x=`midpoint`, y=`ReadPosition`, colour=Sample))+ geom_point(size =
.01)+ ylim(0,300)+ xlim(434000,440000)
Table
DT2 = data.table(
  Sample = Sample,
 Read Midpoint = DT$midpoint,
  Read Position = 1:431001 ,
)
DT2 = data.table(
  Sample = c(Sample),
  endposition1 = 0:431002,
  startposition1 = 0:431002
)
DT$a<-(endpositiongsm1)
DT$b<-startpositiongsm1
endpositiongsm1[,1]
endpositiongsm1<- as.data.table(endpositiongsm1)
(endpositiongsm1)
DT$midpoint <- rowMeans(DT)
DT$
```

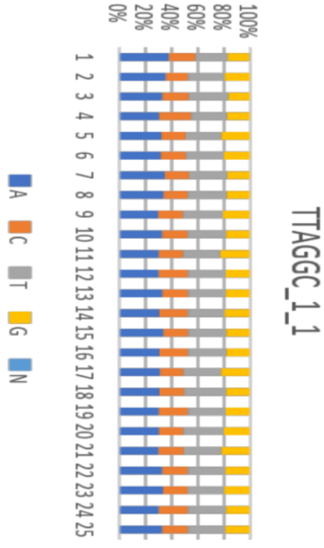
Appendix H: V-Plot Script in Python

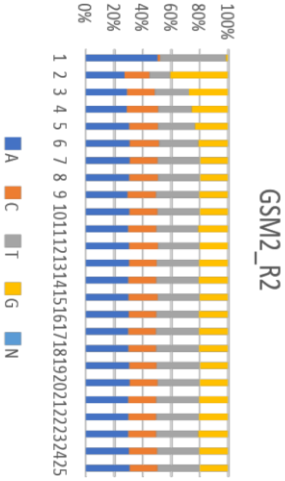
```
import os
import sys
import re
import matplotlib.pyplot as plt
sam_file = open(sys.argv[1])
#input the path of the sam file at the command line
length = 0
midpoint = 0
for line in sam_file:
    if line.startswith("D"):
        split = line.split("\t")
        endposition = (split[7])
        startposition = (split[3])
        midpointnew= (float(endposition)+ float(startposition))/2
        midpoint = str(midpoint) + "," + str(midpointnew)
        newlength = abs(int(split[8]))
        length = str(length) + "," + str(newlength)
length = [ float(x) for x in length.split(',') ]
midpoint = [ float(x) for x in midpoint.split(',') ]
midpoint = midpoint[1:]
length = length[1:]
plt.scatter(midpoint,length, s = .1, alpha = .1)
plt.show()
```

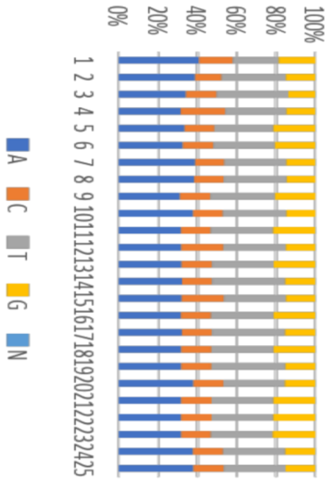
Appendix I: Additional Nucleotide Histograms

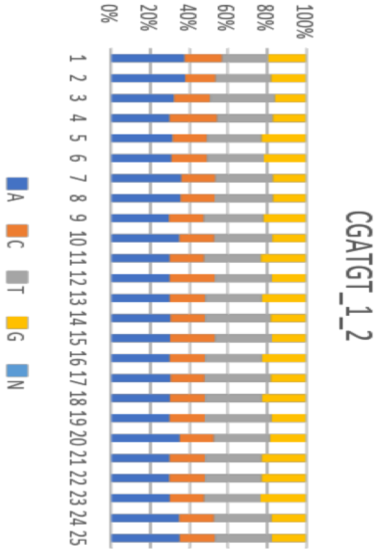

GSM2_R1
A
C
T
G
N

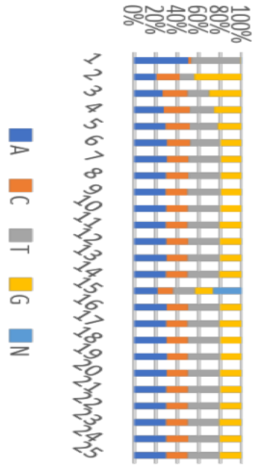

GSM1_R1
A
C
T
G
N

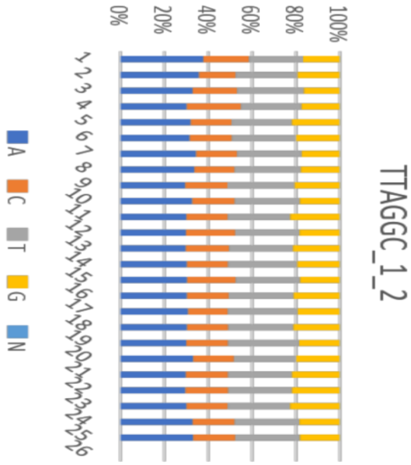

TTAGGC_1_1
A
C
T
G
N

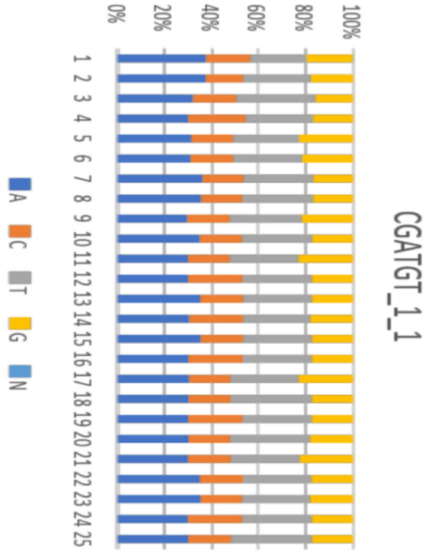

GSM2_R2
A
C
T
G
N

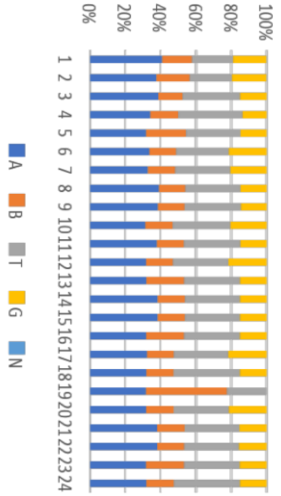

ATCACG_1_2
A
C
T
G
N


CGATGT_1_2
A
C
T
G
N


GSM1_R2
A
C
T
G
N


TTAGGC_1_2
A
C
T
G
N


CGATGT_1_1
A
C
T
G
N


ATCACG_1_1
A
B
T
G
N