2014-06-01

Analysis of Nucleosome Mobility, Fragility, and Recovery: From Embryonic Stem Cells to Invitrosomes

Ashley Nicolle Wright
Brigham Young University - Provo

Analysis of Nucleosome Mobility, Fragility, and Recovery: From Embryonic Stem Cells to Invitrosomes

Ashley N. Wright

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Steven M. Johnson, Chair
K. Scott Weber
Joshua A. Udall

Department of Microbiology and Molecular Biology

Brigham Young University

June 2014

ABSTRACT

Analysis of Nucleosome Mobility, Fragility, and Recovery: From Embryonic Stem Cells to Invitrosomes

Ashley N. Wright
Department of Microbiology and Molecular Biology, BYU
Master of Science

Several factors direct the placement of specific nucleosomes, which in turn have the ability to regulate DNA accessibility. These factors include, but are not limited to, nucleotide sequence preference, nucleotide modifications, the type of histones present within the nucleosome, and the presence of additional transcription factor or chromatin remodelers. A combination of these and other factors are responsible for tightly controlled efficient transcription within the eukaryotic cell. In order to contribute to the understanding of these complicated processes, three separate hypothesis-driven investigations were carried out. First, we looked into nucleosome positioning and phasing within closely related cells lines. Second, we examined domain level nucleosome occupancy on various portions of the chromosome. Finally, we generated a novel method that significantly reduces data loss in *in vitro* nucleosome reconstitution experiments caused by nucleosome fragment-end bias. All three of our investigations yielded separate results. First, by examining positions and phasing patterns within similar cell types we find common patterns and minor differences within similar cell types. The presence of minor differences in nucleosome positions may cause unique expression patterns. Secondly, we found that decreased domain level nucleosome occupancy at the chromosome arms is not caused by the presence of a class of nucleosomes, termed fragile nucleosomes. Finally, we found that when our nucleosome recovery method is applied conservatively to our dataset, it is possible to recover 80% of the lost nucleosome reconstitution data.

Keywords: epigenetic regulation, nucleosome positioning, *in vitro* nucleosome reconsitution

Acknowledgments

It is difficult to appropriately describe all of the guidance and encouragement I have received in only few hundred words. Between learning the new protocols, re-running failed experiments, processing data, and analyzing data I can confidently say I have grown both as a scientist and as a person. All of this could not have happened without the exceptional support system I was fortunate to be surrounded by.

I would like to thank my advisors: Dr. Steven Johnson, for his willingness to answer my continually stream of questions, for his valuable insights, and for his constant guidance throughout the graduate program. My graduate experience would not have been the same without such a positive mentor and for that I will be forever grateful. Dr. Scott Weber for his knowledge of molecular biology and positive encouragement. I truly enjoyed being his teaching assistant. Dr. Joshua Udall for his knowledge of bioinformatics. It is because of him that I was able to acquire my programing experience, which will be valuable for the rest of my career.

Additionally, I would like to thank Paul Bodily and his advisor Dr. Mark Clement whose perl scripts allowed the invitrosome project to be so successful. I also would like to thank them both for teaching me basic perl and bioperl programing. Colton Kempton and the other members of the Johnson lab deserve a thanks for their knowledge, assistance, advice, and love of science. Lastly, I would like to thank my wonderful husband for his endless support and encouragement over the last two years.

Table of Contents

List of Tables

List of Figures

Chapter 1. Background

The most recent published draft of the human genome contains approximately three billion base pairs [1, 2]. This enormous number of nucleotides is not found in one long string, but rather packaged neatly into 23 chromosomes. With the exception of gamete cells (i.e. female ova and male sperm cells), all the cells in our bodies have two copies of the genome, such that each cell contains a total of more than six million base pairs. To put this number into perspective, a single base pair is about 0.34 nanometers long, so six billion base pairs should equal a total length of approximately two meters of DNA. If we estimate the average human body to have about 50 trillion cells, this would suggest that there is a total of 100 trillion meters of DNA within the human body. This may seem meaningless, but consider the fact that the distance between the sun and the earth is about 150 billion meters [3]. This means that the total length of all the DNA molecules in the body are long enough to reach from the surface of our planet to the sun and back more than 300 times! Yet this entire length of our genome is packaged within the microscopic space of the cell's nucleus whose diameter measures in at a mere six micrometers [4]. How is this possible?

The answer lies with a family of proteins called histones. Characterized in 1974, histones are a family of small, positively charged proteins that share a common histone fold [5]. By using the electrostatic force generated by the attraction between the positive histone proteins and the negative phosphate backbone of the DNA molecule, histones have the ability to condense the DNA to a length six to seven times shorter than its original length [6]. This combination of the DNA and histone proteins is called a nucleosome. Nucleosome protein structure consists of an octomer containing two dimers made up of the histones H2A and H2B and a tetramer made up of two of each of H3 and H4 histones [7]. Grooves in the octomer act as a helical ramp around which the target DNA (approximately 147 bp) is wrapped about 1.7 times in a left-handed manner [8]. Additional compaction is generated by a fifth histone protein, H1, that wraps up the intervening DNA, termed linker DNA, bringing neighboring nucleosomes closer together [9]. Further compaction involves additional complex coiling of the DNA molecule until the well-known chromosome structures are achieved.

In addition to facilitating the first step of DNA compaction, nucleosomes are also the first barrier the transcription machinery must overcome if successful gene expression is to take place [10]. To illustrate this concept, consider the initial steps that must be taken before the RNA polymerase can begin transcribing a gene. Almost all those steps involve the interaction between a DNA binding protein (e.g. a transcription factor or an activator) and target sequence found upstream or within the gene itself. The consequences of one or several of those target sites being bound up in a nucleosome would prevent the essential protein-DNA interactions that are required for the initiation of transcription. However, say that this same nucleosome was shifted to the left or right. These essential sequences are now free to interact, allowing transcription initiation to begin. Understanding nucleosome positions and the factors that direct nucleosome positioning is key to our complete understanding of gene regulation within our own bodies.

To illustrate the importance of a proper understanding the process of nucleosome positioning, consider the example of gene therapy. The principle behind this type of therapy is simple; replace the native mutated DNA with a new functional copy [11]. Unfortunately, several problems have arisen as this principle has been put in practice. A common problem with gene therapy is that successive generations are unable express the new functioning copy, despite the fact that the first generation of cells was successfully able to do so. It has been suggested that this is a result of gene silencing caused by nucleosome placement [12]. The new copy lacks the signals that prevent nucleosomes from generating heterochromatin over the new copy, effectively silencing the new gene. Initial studies suggest that if the copy of the DNA could be modified in such a way to prevent this form of epigenetic regulation, this common problem would be solved (unpublished data). This problem illustrates the importance of understanding the process of nucleosome positioning within the cell. Having a more complete understanding of the factors that determine where a nucleosome places itself and how patterns of nucleosome positioning affect gene transcription will allow us to generate medical therapies that could very well allow us to finally overcome diseases that are currently incurable.

Over the last 20 years the list of factors that direct and regulate nucleosome positions has begun to take shape. In general, these factors affect the nucleosome in one of two ways. Either the factor affects energy of the steric hindrance to be overcome in order for the DNA to bend thereby increasing or decreasing the favorability of nucleosome formation or it affects the accessibility of the DNA to the nucleosome [6].

Numerous studies support the notion that nucleotide sequences have a significant effect on nucleosome binding. Whole genome nucleosome mapping approaches show that nucleosomes, in general, prefer GC rich regions and avoid strings of poly-A/Ts [6, 13]. Several studies measuring histone-binding affinities have found that high-affinity sequences generally contain 10-11 periodic dinucleotide motifs [14, 15]. It is thought that this periodicity helps to form and stabilize a favorable confirmation by minimizing the costs of DNA bending [12]. Specifically, there is a preference for G/C to be placed where the major groove is compressed and A/T where the minor groove is compressed.

DNA methylation specifically has been shown to have a direct effect on nucleosome positioning. CpG methylation creates sites were the DNA becomes stiffer than the surrounding DNA [16]. This creates instability if the sequence is bound up by a nucleosome, especially so if the methylation is in a place where it would come between the histone core and the DNA [17]. By generating unfavorable sites, methylation can change individual positions as well as the positions of large groups of neighboring nucleosomes [17]. This in turn can change the local DNA accessibility and gene functionality. However, a very recent 2012 study has shown that when methylated DNA is used in *in vitro* reconstitutions, the nucleosomes formed are actually very stable [18]. While these studies contradict one another, it is clear DNA methylation affects nucleosome formation and positioning.

The placement of a nucleosome can also be dictated by the presence of histone variants within the octomer. It is only recently that histone variants have been a subject of interest. H3.3 and H2A.Z are two variants that have been studied extensively. These two are histone isoforms of H3 and H2A, respectively. H3.3 is thought to function in regions where histones have been displaced as a result of chromatin remodelers involved in transcription [19]. H2A.Z is controversial, as contradictory roles have been observed. In some cases, H2A.Z has been shown to be involved in histones that need to be quickly displaced, such a promoter regions [20]. In other cases, this histone variant has been seen to play a part in histones that occupy and are responsible for chromatin silencing [21]. In either variants case, it is clear that the distribution of this histone is non-random. H2A.Z and H3.3 both seem to play a part in histones that place themselves in positions that have a distinct purposes, such as quick removal to allow quick transcription or to mark a section for silencing.

Finally, steric hindrance caused by the presence of a bound protein can also affect nucleosome positioning. A bound transcription factor, polymerase, or other DNA binding protein can position a nucleosome by sequestering portions of the DNA that favor nucleosome formation [22, 23]. This in turn forces the nucleosome to seek an alternative site on which to form. It has been observed that nucleosomes near transcription start sites often position themselves directly adjacent to a bound protein, which in turn can cause other local nucleosomes to distribute themselves evenly beginning with this first nucleosome in a phenomena know as nucleosome phasing [15, 24].

A combination of all of these factors is likely what controls transcription. While each different variable has been shown to have an effect on nucleosome positions, one of these variables alone cannot account for the complex and dynamic patterns of positioning observed in multiple organisms [6]. A more complete understanding of the dynamic processes that control these patterns and in turn regulate transcription will provide enormous insight in many fields including development, evolutionary and molecular biology.

In order to make a significant contribution to the understanding of these chromatin processes, I have performed three hypothesis-driven sets of experiments that: 1) analyze differences in nucleosome position and phasing between closely related cell lines on the local scale; 2) examine domain-level differences in nucleosome occupancy on various portions of chromosomes; and 3) test the validity of *in vitro* nucleosome reconstitution techniques in common use in the nucleosome positioning field.

Chapter 2. Nucleosome Mobility and Cellular Differentiation

2.1 Introduction

Embryonic stem (ES) cells represent a major leap forward in biology and medicine. Their discovery has allowed for 1) the creation of *in vitro* models of early mammalian development; 2) the generation of animal models of that accurately represent the progression of human diseases; and 3) the production of biologically identical differentiated cell types for cell replacement therapy [25]. However, when the factors that maintain their totipotent state are removed, ES cells undergo differentiation [26, 27, 28]. If the appropriate combination of factors is added to the cells at this point, ES cells will take on the characteristics of a specific embryonic germ layer (e.g. mesoderm, endoderm, or ectoderm)[25, 29]. Human embryonic cells are unique in that when induced with BMP4, the cells will take on the characteristics of the trophectoderm, which is responsible for the generation of the placental interface between mother and embryo [30]. Other eukaryotic wild-type ES cell populations, such as mouse ES cells will not do this; they are only capable of reflecting the potential of the normal origin of embryonic cells, the inner cell mass.

It has been observed that ES cell lines, when allowed to differentiate on their own and no additional factors are added, will choose a specific cell type a majority of the time. The cell type into which it develops varies from line to line [31]. We proposed that this preferential differentiation of ES cells may be a consequence of differential transcription factor bind that will be easily detected by nucleosome phasing. As described above, the phenomena of phasing is caused by the presence of some type of barrier (e.g. a bound transcription factor). As it occupies a prime position, this barrier forces local nucleosomes to positions directly adjacent to this barrier. This has a cascading effect on the neighboring nucleosomes, which respond by positioning themselves at precise intervals from the nucleosomes directly adjacent to the barrier [24]. This creates a distinct pattern of nucleosomes that can be clearly detected in nucleosome mapping experiments.

We hypothesize that this is the result of differences in transcription regulation. It stands to reason that the differences in cell lines are not due to the cell's genomic sequence. While each cell line's origin is different, it is unlikely that minor variation between the genomes could cause such a drastic difference [31]. This leaves transcription regulation as the most likely cause of these cell line preferences. However, detecting these differences will be very difficult as they are likely to be very small subset of transcription factors relative to general regulation by transcription factors going on within these ES cells.

The purpose of this investigation is to test the hypothesis that minor motif level differences in nucleosome positions exist between cell lines, and can reveal the different transcription factors binding patterns that may be causing the differentiation preference. A simple yet novel way to detect these minor differences will be through the use of high throughput sequencing. As mentioned previously, bound proteins are very often the cause of nucleosome phasing. Ideally, by sequencing nucleosome fragments and mapping them back to the genome, we will be able to detect phasing in the various cell lines. Nucleosome phasing unique to each of the cell lines will act as a putative yet easily detectable marker for transcription factors and other bound proteins unique to this cell line. We were able to identify cell-line specific differential sites of these bound proteins. Though this project was put on hold, we will be able to identify cell-specific transcription regulation occurring in the different stem cell lines. We will also be able to provide identification of the bound factors by completing additional experiments isolating bound factors at sites of identified phasing. This information will further our limited understanding of stem cell differentiation, which has already been shown to be extremely powerful tool.

2.2 Materials and Methods

Isolation of mononucleosome core DNA fragments

The following is modified from [15]. Flash frozen H9 hES cells were ground to a fine powder in liquid nitrogen using a mortar and pestle. An equal volume of 0.34 M sucrose/Buffer A (15 mM Tris-HCl at pH 7.4, 15 mM NaCl, 1 mM DTT, 60 mM KCl, 0.5 mM spermidine, 0.15 mM spermine, 25 mM bisulfite) was added to powered cells. After thawing on ice, $CaCl_2$ and micrococcal nuclease (Roche) resuspended at 300 U/μL were added for final concentrations of 1 mM and 25 U/μL, respectively, followed by incubation at 16°C for 12 min to liberate the mononucleosome cores. The reaction was stopped by the addition of an equal volume of worm lysis buffer (0.1 M Tris-HCl at pH 8.5, 0.1 M NaCl, 50 mM EDTA, 1% SDS), and proteins were removed by treating with one-tenth volume proteinase K (20 mg/mL in TE at pH 7.4) for 45 min at 65°C, followed by phenol, phenol/chloroform, and chloroform extractions and ethanol precipitation. After RNase treatment and phenol/chloroform, chloroform extraction, separation of the micrococcal nuclease-digested DNA into mono-, di-, tri-, and multinucleosome DNAs was done using a 4% UltraPure Agarose (Invitrogen) gel run at 100 V for 4 hours. DNA from the mononucleosome DNA band was extracted from the gel using the QIAquick Gel Extraction Kit (Qiagen) following the standard protocol, with the exception of allowing the isolated gel sample to incubate in Buffer QG at room temperature until dissolved. Once isolated, the concentrations of each fragment was quantified using Agilent bioanalyzer high sensitivity DNA analysis kit in the BYU DNA sequencing center.

Library Preparation and Illumina Sequencing

Library preparation of isolated sequences was completed at the USC sequencing facility. The facilities standard protocol can be found on their website (epigenome.usc.edu). The ends of fragments were repaired using the Epicenter Biotechnology End-Repair Kit (EpiBio). Sequence fragments were combined with 10X end repair buffer, dNTPs, ATP, end repair enzyme in a 50 μL reaction. The reaction was incubated at room temperature for 45 min. The repair enzyme was deactivated by incubating the reaction at 70°C for 10 min. The reaction was cleaned by using a standardized magnetic bead cleanup protocol. 1.5x volume of beads were added to the reaction and incubated for 5 min. A magnet was used to migrate the beads, allowing the solution to clear. Supernatant was discarded and the beads were washed several times with 70% ethanol. Ethanol was removed and 42 μL distilled water added. Supernatant was pipetted and used in the A-tailing reaction. A-tailing ligation was prepared using NEBNext dA-tailing Module (NEB). To the eluted end-repaired template NEB Next d-A buffer and Klenow fragment was added. The reaction was incubated at 37°C for 30 min. The reaction was cleaned up using the same magnetic bead protocol described above. The enzymatic adapter ligation kit (Enzymatic) was used to ligate adapters to the eluted reaction. To the eluted A-tailed template 2x rapid ligation buffer, stock adapters, and T4 ligase enzyme were added. The reaction was incubated at room temperate for 10 min and cleaned up using the same magnetic bead clean up protocol. The ligated template was eluted in 16 μL of elution buffer. The concentration of the template was quantified using the NanoDrop 2000 UV-Vis Spectrophotometer. The adapter ligated templates were amplified using PCR amplification to produce enough product in preparation for cluster formation and sequencing. Illumina sequencing was preformed at the USC sequencing facility using the Illumina Hi-Seq2000 system.

Bioinformatic analysis using NuMap

Once sequenced the collected fragments were processed using NuMap. NuMap is a set of computational programs developed by our collaborator Anton Valouev. The purpose of this set of programs is the analysis of nucleosome mapping data (MNase-Seq), and histone modification data. To access the program specifications and a user guide please visit http://www-hsc.usc.edu/valouev/NuMap/NuMap.html. This program first calls all of the positions of the nucleosomes within the cell. Once completed, called positions are used to detect phasing patterns.

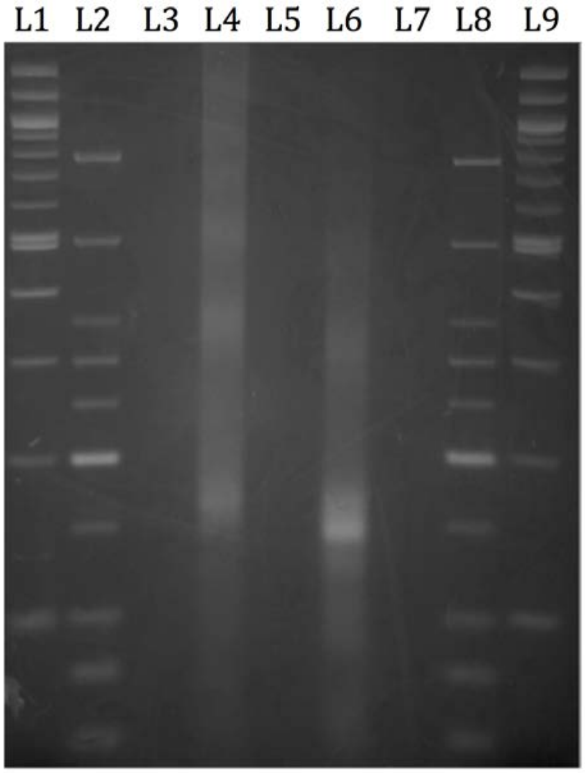


Figure 2.1: The digestion gel shows a the muti-, tri-, di-, and mononucleosome bands of the 100 U/200 μL digestion (lane 4) and the 5,000 U/ 200μL digestion (lane 6). For point of reference, ladders were run along side of the digestion. These include a 100 bp ladder (lanes 1 and 8) and a 50 bp ladder (lanes 2 and 9)

2.3 Results

In order to determine phasing patterns, nucleosome fragments were isolated and mapped back to the genome. The results of the MNase digestion are shown in Figure 2.1. Two separate digestions were preformed. The digestion on the left (lane 4 of Figure 2.1) shows clear multi-, tri-, di-, mono-, and sub mononucleosomal bands. However, the greater intensity of the mononucleosomal band of the digestion on the right (lane 6 of Figure 2.1 suggests a more complete digestion. The mononucleosome bands from both digestions were dissected from the gel. A size selection step was necessary for the sequencing to contain exclusively mononucleosomes, and avoid di-, tri-, and higher order nucleosome structures.

The resulting sequenced reads were analyzed by our collaborator, Dr. Anton Valouev. Phasing patterns were detected that conformed to observed patterns in the literature [32].

2.4 Future Directions

This project was unfortunately put on hold. It was decided that further investigation was need before continuing with other cell lines. While the phasing patterns did indicate the presence of transcription factors, it was decided that the transcription factors should be directly isolated to provide further confidence. Until this additional investigation has been completed, this project will continue to be put on hold.

When the project is ready to continue, additional hES cell lines will be digested following the same protocol described above allowing for the isolation of nucleosome fragments. The nucleosome positions provided by these additional digestions should allow us to detect any minor differences in cell lines as indicated by specific phasing patterns. Nucleosome phasing unique to the cell lines will act as a detectable marker for transcription factors and other bound proteins unique to this cell line. Through identification of cell-line specific local phasing patterns in combination with the isolated transcription factors bound at these same sites, we should be able to identify cell-specific transcription regulation occurring in the different stem cell lines.

Chapter 3. Localized Fragile Nucleosome Bias at Chromosome Ends

3.1 Introduction

As mentioned previously, it has been established that nucleotide sequences affects nucleosome binding. However, the actual strength of sequence influence is still up for debate. The energy required to move a nucleosome is relatively small compared to other processes that take place during transcription [13, 33]. Often the transcription machinery contains or is capable of recruiting chromatin-remodeling enzymes [34]. Comparative studies of chromatin formation *in vivo* and *in vitro* (performed in *Saccharomyces cerevisae*) show that while the bulk of the nucleosome do not seem to be influenced by sequence preference, there are a subset of nucleosomes that are predictably positioned by the DNA sequence [33]. It is unclear, however, whether the same is true within more complex organisms that have multiple cell types and nucleosome positions more strongly affected by chromatin remodeling.

To address these issues, a recent 2013 study was performed using the nematode worm *Caenorhabditis elegans* [14]. Through salt dialysis, nucleosomes were formed on naked worm DNA. High-throughput sequencing was used to create genome-scale maps of these *in vitro* reconstituted nucleosome positions. These were then compared to two different *in vivo* maps generated using whole worms (Illumina and SOLID sequencing). Additionally, the *in vitro* maps were compared to maps generated from *C. elegans*, embryonic tissues, adult somatic cells, and germ cells (Illumina sequencing). This extensive comparison to invitrosomes (*in vitro* reconstituted nucleosomes) allowed for the systematic identification of sequence-determined positions from those positioned by other external factors within a multi-cellular organism.

One of the most interesting results of this study was the nucleosome distribution across the chromosomes. The *in vitro* data clearly shows that across the autosomes, the nucleosomes distribute very evenly. However, the *in vivo* generated data did not show this same even distribution. This data, shown in Figure 3.1 reveals that there is an uneven distribution of nucleosomes on the chromosome arms, a relative depletion, right before the telomeric regions. We believe two possibilities exist to explain this distribution. First, there are simply fewer nucleosomes on the chromosome arms. Secondly, it is possible that there are a high number of fragile nucleosomes (i.e., nucleosomes that are highly susceptible to Micrococcal nuclease digestion) within these regions.

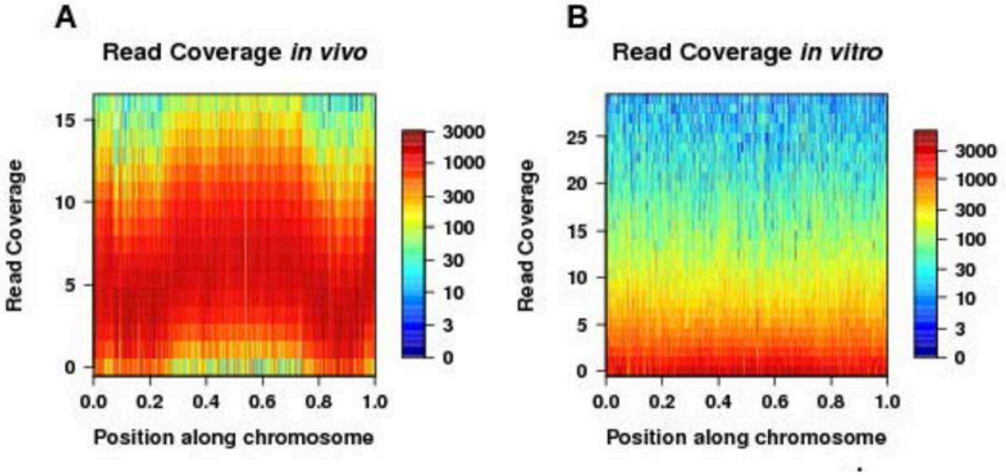


Figure 3.1: *In vivo* chromosomal nucleosome distribution verses *in vitro* chromosomal nucleosome distribution. While distribution is relatively even across the *in vitro* generated dataset (right), this is clearly not true of the *in vivo* generate set (on left). In the *in vivo* set, there is a clear depletion of nucleosomes at the chromosome arms and an increase at the chromosome center. Figure modified from [14].

Studies in yeast have shown that histone composition and the conformational state of the nucleosome may alter the relative Micrococcal nuclease (MNase) resistance of an individual nucleosome. For example, it has been documented that nucleosomes containing H2A.Z and H3.3 isoforms are unstable and appear hypersensitive to MNase resulting in the loss of their associated DNA during a typical MNase digestion [21]. Additionally, a portion of the core DNA may become spontaneously unwrapped from the histone core. If this confirmation were stabilized through adjacent protein interactions, this would render the nucleosome less resistant to MNase digestion. Several lines of evidence support the notion that variable MNase resistance does occur (termed nucleosome fragility) throughout the genome [23, 22, 35].

Fragile nucleosomes can be identified by their higher abundance in incomplete MNase digests relative to complete digests [35]. To identify fragile nucleosomes, chromatin is subjected to MNase digestion without formaldehyde cross-linking. As shown in Figure 3.2, mononucleosomal DNA is recovered at two separate time points during a digestion: at a point of incomplete digestion and at the point of complete digestion (i.e. at which time all present chromatin has been reduced to mononucleosomes). These samples are sequenced and the nucleosomal positions and occupancy at those positions are determined. Examining nucleosome maps generated from these two conditions allow fragile nucleosomes to be identified, as they will be abundant in the incomplete digestions but mostly absent in the complete digestion [35]. Fragile nucleosomes can be rescued using crosslinking. Studies in yeast suggest that fragile nucleosome distribution is not random and is likely due to physiological conditions [35].

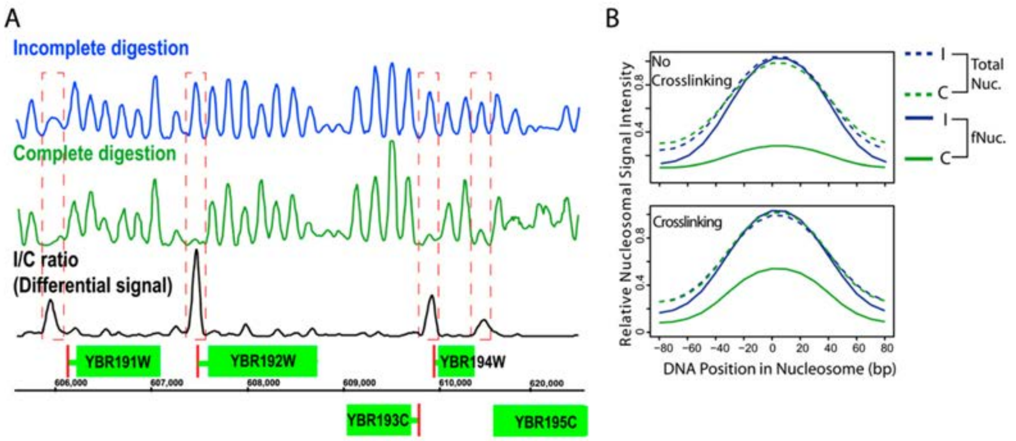


Figure 3.2: A. Fragile nucleosomes are defined by their presence in incomplete digestions and absence in complete digestions (represented by peaks within the red broken lines). B. Additionally, fragile nucleosomes can be partially rescued using formaldehyde cross linking. Figure modified from [35]

The purpose of this investigation is to test the hypothesis that there is a high number of fragile nucleosome localized to the chromosome ends within the *C. elegans* genome. To detect the presence of these nucleosomes, a series of digestions were preformed using variable concentrations of MNase as well as varying lengths of digestion time. Fragile nucleosomes are highly sensitive to digestion and so should be the first nucleosomes to be liberated by and the first to be lost withe further digestion [35]. Therefore, a majority of fragments isolated from the incomplete digestion conditions (e.g. short digestion time and small concentration of Mnase) should represent the positions of fragile nucleosomes. Incomplete digestions used to define these fragile nucleosomes will be referred to light digestions. Complete digestions will be referred to as heavy digestions. In order to support the proposed hypothesis, these lightly digested fragments should also map back to the sites of decrease nucleosome occupancy seen in the *in vivo* maps. Accordingly, a majority of the heavily digested fragments should map to regions outside of these regions of decreased occupancy. These results should shed some light on this unique domain-level chromatin architecture presence in the *C.elegans* genome.

3.2 Materials and Methods

Variable MNase Digestion

The following is modified from [10]. Mixed stage, wild-type (N2) *C. elegans* were cultured on DH5alpha E. coli, flash frozen with liquid nitrogen in 0.34 M sucrose/Buffer A (15 mM Tris-HCl at pH 7.4, 15 mM NaCl, 1 mM DTT, 60 mM KCl, 0.5 mM spermidine, 0.15 mM spermine, 25 mM bisulfite), and ground to a fine powder in liquid nitrogen using a mortar and pestle. After thawing on ice, $CaCl_2$ and micrococcal nuclease (Roche). Separate digestions were performed using increasing units of micrococcal nuclease (40, 160, or 640 U/μL) which were incubation at 25°C for 2, 6, 18, or 54 min to liberate the mononucleosome cores. The reaction was stopped by the addition of an equal volume of worm lysis buffer (0.1 M Tris-HCl at pH 8.5, 0.1 M NaCl, 50 mM EDTA, 1% SDS), and proteins were removed by treating with one-tenth volume proteinase K (20 mg/mL in TE at pH 7.4) for 45 min at 65°C, followed by phenol, phenol/chloroform, and chloroform extractions and ethanol precipitation. After RNase treatment and phenol/chloroform, chloroform extraction, separation of the micrococcal nuclease-digested DNA into mono-, di-, tri-, and multinucleosome DNAs was done using a 2% UltraPure Agarose (Invitrogen) gel run at 50V for 4 hours, and DNA from the mononucleosome DNA bands were extracted from the gel using the QIAquick Gel Extraction Kit (Qiagen) following the standard protocol, with the exception of allowing the isolated gel sample to incubate in Buffer QG at room temperature until dissolved. Once isolated, the concentrations of each fragment was quantified using the NanoDrop 2000 UV-Vis Spectrophotometer. Series of digestions allows us to sample from a variety of conditions that contrast one another. The ten digestion conditions used are described in Table 3.1 and pictured in Figure 3.3 below.

Table 3.1: Table shows the various digestion conditions and the corresponding barcode. Additionally, the table indicates whether the digestion condition was designated as a heavy or light digestion. Finally, the lane from figure 3.3 representing the individual digestion is indicated

Digestion Time (min)	**U/μL**	**Barcode**	**Digestion Designation**	**Lane**
2	40	TTGT	Light	3
2	160	ACGT	Light	4
2	640	CAGT	Light	5
6	40	GCTC	Light	7
6	160	TGCT	Light	8
6	640	CCCT	Heavy	9
18	40	AACT	Light	11
18	160	GCAT	Heavy	12
54	40	CGAT	Heavy	15
54	160	GGGT	Heavy	16
54	640	TAAT	Heavy	17

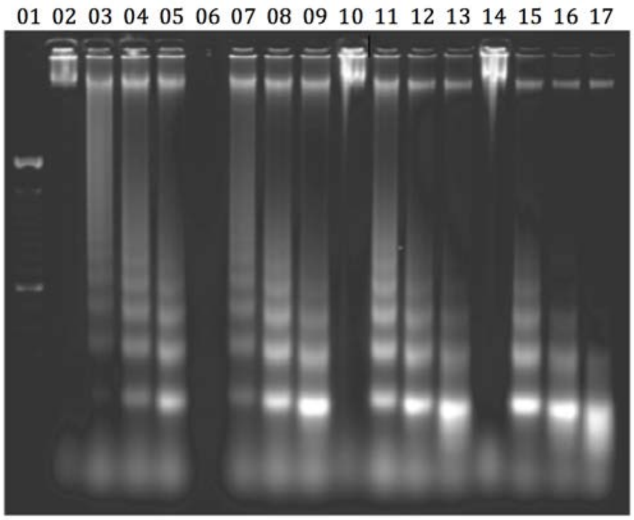


Figure 3.3: Gel shows digestions performed with increasing concentrations of micrococcal nuclease (0, 40, 160, and 640 units) for varying lengths of time (2, 6, 18, 54) at 25°C. Digestions used in this study are listed in Table 3.1. Figure modified from [12]

End Polishing and Barcode Ligation

Fragments ends were polished in preparation for barcode ligation. Mononucleosome fragments were combined with 10 μL 10x T4 DNA ligase buffer, 5 μL T4 DNA polynucleotide Kinase, and 55 μL of water to produce 100 μL reaction. This was run at 37°C for 3 hours. The reaction was quenched by adding 200 μL of 1x stop buffer (0.2% solution SDS) and 1 μL of glycogen. 1 mL of Ethanol was added and the contents placed at -80°C overnight. The following morning the contents were spun to pellet the contents followed by a 500 μL ethanol wash. The ethanol was removed and the resulting pellets resuspended in 33 μL of TE. Polishing continued with the addition of 4 μL NEBuffer, 2 μL dNTPs, and 1 μL of T4 polymerase. The reaction was carried out at 12°C for 15 min. The reaction was quenched with 200 μL stop buffer. The remaining dNTPs were removed using Qiagen PCR purification kit. The polishing process was completed by combining the eluted 32 μL of DNA with 5 μL of NEBuffer 2, 10 μL of 1mM ATP, and 3 μL of Klenow exonuclease. This reaction was run for 30 min at 37°C. The polished sequences were separated from the other contents of the reaction using Qiagen PCR kit. Isolated polished fragments were then ligated to an adapter to facilitate barcode ligation. Barcode ligation followed. Once the fragments were isolated post-digestion, they were ligated to barcodes. Doing so allowed us to pool our fragments together for sequencing and then easily sort sequenced reads based on digestion condition afterwards. The barcodes used are listed in table 3.1 above. Barcodes were designed so that the likelihood of finding the equivalent sequence at random within a sequence would be minimal. Final concentrations of the libraries were quantified using Agilent bioanlyzer high sensitivity DNA analysis kit in the BYU sequencing center. Portions from each of the libraries were combined into a single sample in preparation for sequencing. The amount each of the libraries contributed to the final was based on bioanalyzer concentration, such that in theory each digestion was represented equally within the combined sample. In reality, a single condition (lane 16 of Figure 3.3) was over represented and made up the vast majority of our sequenced reads.

High-Throughput Sequencing

Our pooled samples were sent to the USC sequencing facility for sequencing. A single lane of Illumina sequencing was preformed at the USC sequencing facility using the Illumina Hi-Seq2000 system. Due to the over representation of one of our conditions, phiX DNA was added to our sample to increase the diversity of the first base reads and facilitate sequencing.

Bioinformatic Analysis

The resulting reads were first parsed into separate files based on the identifying barcode. The parsed reads were then each mapped to a template file contain the WS190 Celegans genome. Alignments were generated using the program BLAT. Parameters were set so that reads only mapped with sequences that would generate very high alignment scores (maximum of 1 mismatch). All other parameters for BLAT were set to default settings. Multiple alignments were processed using a perl script written by Fredrick Tan, such only the match with the greatest alignment score was retained. The remaining processed and mapped reads were and then sorted based on chromosome of origin. The five autosomes were divided in 10 equal pieces and the number of reads per chromosomal fragment was calculated.

3.3 Results and Discussion

Each chromosome was divided into 10 equal sized increments. For each digestion condition, the number of reads that mapped to each of these individual increments were separated out and summed. Additionally, these numbers were divided by the total number of reads mapping to the individual chromosome within the individual digestion condition. All of these analyses for the individual digestion conditions can be found in appendix A. To identify general trends within the light digestions, the total reads per increment for each light digestion were summed together to determined the summed read distribution (Figure 3.4). These sum totals were added together to determine the total number of reads mapping to each increment originating from all light digestions. These sums were used to determine the percent of reads that mapped to each chromosomal increment within all light digestions (Figure 3.5).This same analysis was completed for the heavy digestions (Figures 3.4 and 3.5).

Examining the summed total read distribution for both light and heavy conditions show a few similar general trends. The summed total distribution for the light conditions shows that across chromosomes I,II, and III there are about the same number of reads distributed evenly within each increment. However, for chromosomes IV and V the number of reads varies greatly, with the greatest number of reads in the center of the chromosome. These trends are illustrated in Figure 3.4). In the summed read distribution for the heavy conditions we see similar trends. Again, across chromosomes I, II, and III we see an equal number of reads across each increment. Additionally, the number of reads on chromosome IV and V varies, through the variability is less dramatic than that is seen within the distributiond for the light conditions. The greatest number of reads for the heavy conditions seems to fall within the center of these two chromosomes. All of these trends are illustrated in Figure 3.4. At first glance, the two summed distributions look different. However, the trends followed within both distributions do not indicate dramatic differences that would support our hypothesis.

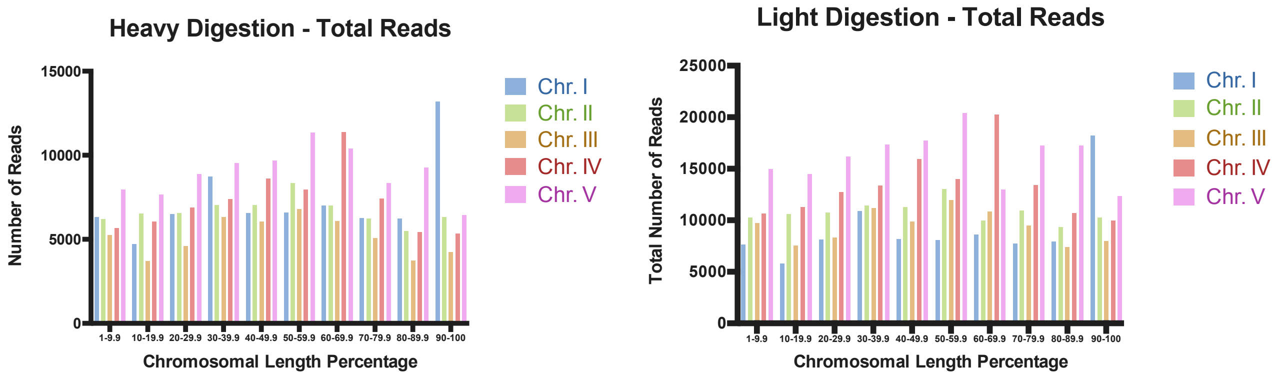


Figure 3.4: Graphs show the total reads mapped to each tenth of the five *C. elegans* autosomes for all heavy or light digestions.

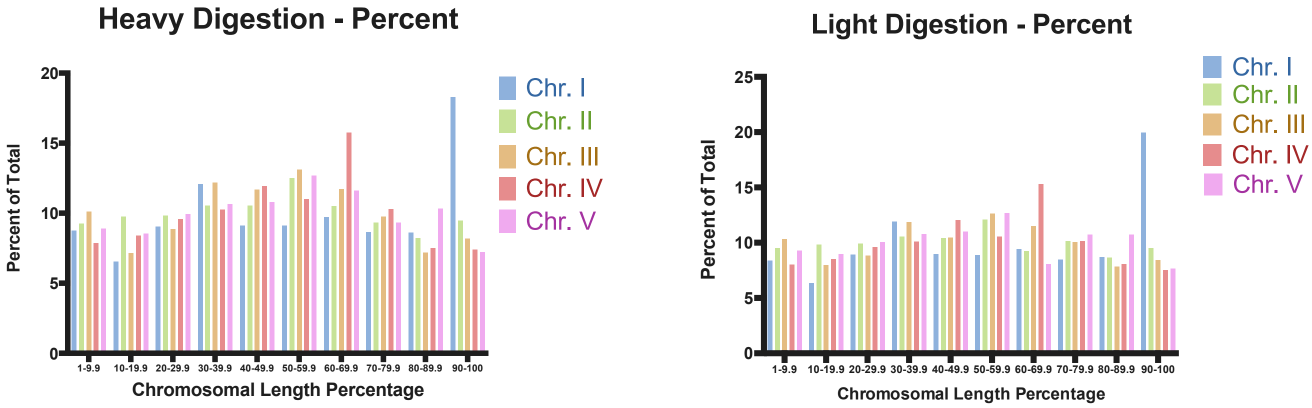


Figure 3.5: Graphs show the total percentage of reads mapped to each tenth of the five *C. elegans* autosomes for all heavy or light digestions

The percentage of reads distributing to each increment for all heavy or all light conditions also indicate similar general trends. An initial glance at the results of these two analysis show far similar distribution relative to the summed distributions. For the light conditions, the percentage of reads across all five chromosomes distribute evenly at or around 10%. Central increments show a slightly greater number of distributed reads, while the outer increments have slightly fewer number of distributed reads. These trends are illustrated in Figure 3.5. Two clear outliers are present. First, there is a clear increase in the number of reads distributed within the final increment of chromosome I. This is an artifact of the collapse of the 5s rRNA genes presence within the WS190 genome in this increment. There also exists a increase in the seventh increment of the fourth chromosome. We are unsure of its cause at this time. Similar trends are observed with the heavy digestions. Very similar trends are observed within percent read distributions for the heavy conditions. In general, the percentage of reads distributing within each increment across all chromosome falls between 10% and 12%. This is illustrated in Figure 3.5. The same two outliers are present within the heavy conditions. We believe the outlier present on chromosome I, is again an artifact of the same genome assembly. The outlier present on chromosome IV, while less dramatic than the one observed within the light conditions is present in the same increment and so it is possible that this two is an artifact of the informatics used.

To further assess the differences between two sets of conditions, we subtracted the total percentage for the light conditions from those of the heavy conditions. This simple analysis clearly shows that little difference exists between the two digestions sets (Figure 3.6). Of note is the clear outlier cause by far more heavy reads distributing to the seventh increment of chromosome V. We are unsure as to what caused this, though one possibility is that this is simply caused by randomness within this particular set of digestions. The remaining large differences seem distributed at random though it should be noted that increases in light read seem to occur at the ends of the chromosomes. Additionally, any relative increase in distributions of heavy reads seem to be focused at the chromosome center. While this would seem to support our hypothesis, further investigation indicates that these increases are caused by very large differences within a single digestion condition, skewing the results.

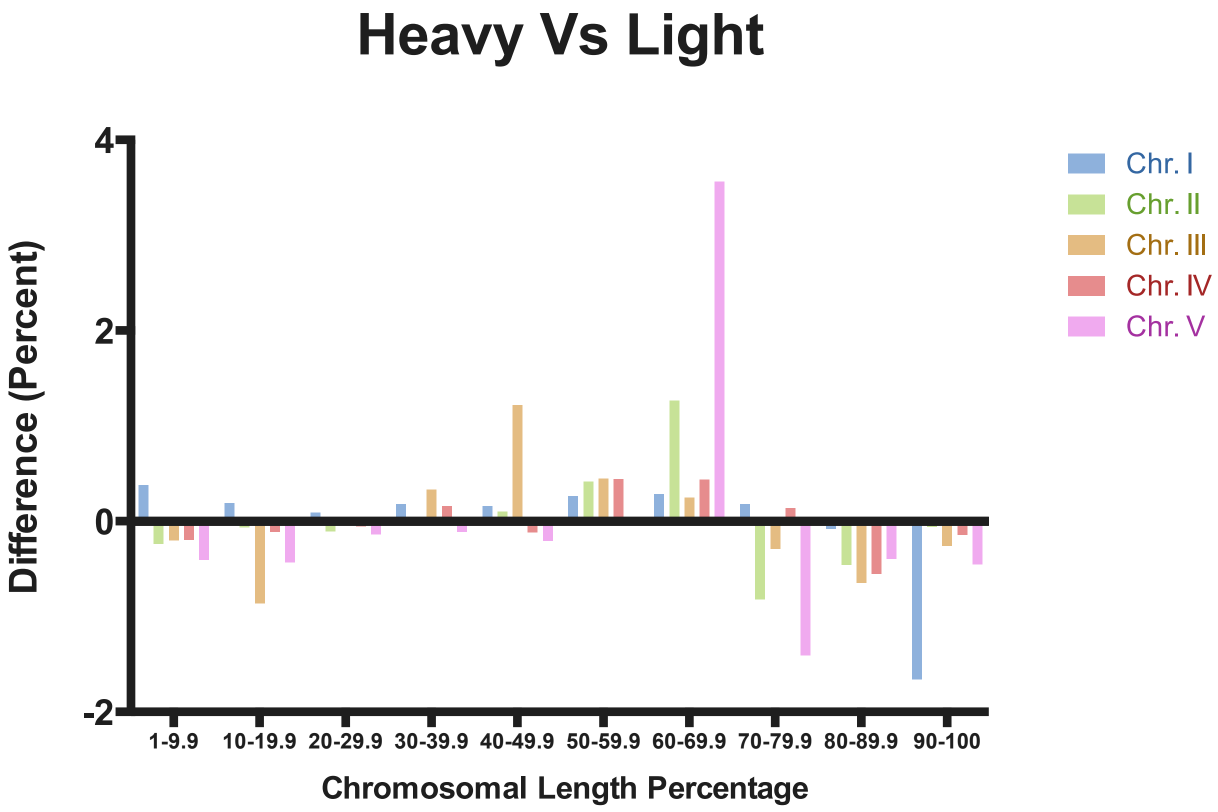


Figure 3.6: Graph shows the difference between condition trends of total percentage of reads mapped to each one-tenth of the five *C. elegans* autosomes.

These results do not support the original hypothesis. We predicted that we would see a far larger increase of nucleosome fragments on the outer ends of the chromosomes with the lighter digestions, while the heavier digestions would produce the opposite pattern (i.e. larger decrease at chromosomes ends). This would be indicative of the presence of fragile nucleosomes concentrated at the chromosome ends. Instead in both cases we see a generally even distribution of nucleosome fragments across all five of the chromosomes. This pattern also seems to contradict the alternative hypothesis proposed; the observed depletion is due to a general depletion of all nucleosomes at the ends of the chromosomes. If this were the case, we would expect to see that same terminal depletion in both the heavy and light digests. As the results fail to indicate any clear regions of significant depletion in either condition, this alternative hypothesis is not supported.

3.4 Future Direction

While these results do not support the proposed hypothesis, further investigation is need before this project is abandoned. The simple analysis described here does not seem to indicate the presence of fragile nucleosome. However, a more complex analysis that mimics the analysis seen in Figure 3.1 needs to be done before we are confident in this conclusion. This will involve dividing the information into much smaller informatics bins. Doing so will allow us to generate results using the same points of reference used in the locke analysis. These result will provide clear indication of the presence or absence of fragile nucleosomes. To provide further confidence, this investigation should be replicated using a dataset that has been subject to initial cross linking. Fragile nucleosomes are rescued when cross-linking is performed before digestions. If fragile nucleosomes are present, we should see the recovery of depletions present in results of this investigation. The results of these two combined investigations should provide enough relevant conclusions for a complete publication.

Chapter 4. Efficient Recovery of Lost Invitrosomes through Comparative Defined-End Analysis

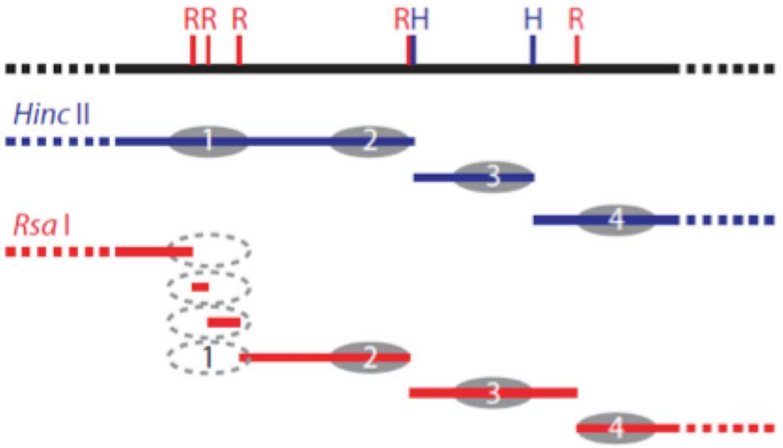
4.1 Introduction

Access to the nucleotide sequence by trans-acting factors is primarily determined by nucleosome positions taken on within the immediate chromatin architecture. As mentioned previously, several factors have been shown to direct and regulate nucleosome positions, including the underlying nucleotide sequence itself [14, 15]. A commonly used method for determining high affinity DNA sequences is through the use of *in vitro* nucleosome reconstitutions. Whole genome applications of this method begin with isolation of naked genomic DNA followed by generation of smaller DNA fragments primarily through sonic sheering or restriction enzyme digestion of the high molecular weight DNA. Recombinant or isolated histone octamers and DNA fragments are then added together in high-salt solution in a stoichiometric ratio such that on average a single nucleosome will form on each individual fragment. The salts in the solution are then dialyzed away, allowing the formation of nucleosomes [36]. The *in vitro* reconstituted assemblies can be compared to their *in vivo* genomic equivalents, allowing for not only the identification of high affinity sequences determined exclusively by their intrinsic DNA sequences, but also the amount of *in vivo* remodeling that occurs within individual cell or tissue types. Such an approach was used by Locke et al. to demonstrate the extent of nucleosome remodeling that happens *in vivo* to the *C. elegans* genome [14].

While *in vitro* nucleosome reconstitutions provide valuable information, the technique contains certain inherent biases that must be overcome before the resulting data is useful. It has been demonstrated that DNA fragment ends can influence nucleosome formation so as to encourage end-proximal nucleosome formation relative to the remainder of the fragment [37, 38]. This preference is termed fragment end bias and can introduce a major hurdle when attempting to identify high affinity sequences as it becomes impossible to determine if *in* *vitro* nucleosome (invitrosome) formation was due to fragment end bias or an actual affinity for the underlying nucleotide sequence. It is thought that this end bias can be overcome by using sonication to generate the needed DNA fragments. In theory, if DNA fragmentation by sonication is random, any fragment end bias generated during nucleosome reconstitutions should be compensated for by the presence of excess random fragment ends which would be evenly spread out over the entire sample and thus produce an uniform background coverage that could be discounted. However, sonication is not a completely random process. Sonication more commonly generates fragment ends within sequences containing runs of polyAs or polyTs. Thus sonication may not be a solution to the fragment end bias dilemma [39, 40]. One option to overcome this hurdle is to discarding nucleosome positions that fall near fragment ends. This is only an option if the ends of the DNA fragments that are being used in the reconstitution experiments are known. Even if this is the case, the amount of data discarded using this option is often a large portion of the potentially meaningful data. This presents a major limitation to nucleosome reconstitution, as it requires an excessive amount of time and materials to guarantee enough usable data is generated once end-proximal nucleosome positions are discarded. Such an approach to overcome potential end-bias was used by Locke et al. in their analysis [14]. In the following investigation, we propose to valid our novel approach for addressing fragment end bias that eliminates the need of discarding large portions of the data produced in these type of experiments. We apply our approach to the Locke et al. data set and show that we can recover up to 80% of the discarded data.

4.2 Approach

Currently using conventional approaches, two classes of DNA loci are typically excluded from invitrosome analysis or have invitrosomes discarded in order to eliminate potential end bias. When DNA fragment ends are defined, 1) any invitrosome found to map within a defined number of nucleotides from a fragment end is classified as suspect of fragment end bias and is discarded. 2) DNA fragments digested to sizes too small for reconstitution (less than 147 bp) are lost from invitrosome analyses [14]. We propose that both of these classes of excluded loci can potentially be recovered and analyzed by performing nucleosome reconstitutions on two DNA samples digested by two different restriction endonucleases. Our approach is such that each individually digested DNA sample is used for separate nucleosome reconstitutions and then invitrosome positions from the two experiments are identified by mapping sequenced mononucleosomal core DNAs back to the original source of DNA. For each individual reconstitution experiment, invitrosomes that may suffer from end-effect bias can be identified by defining a specific number of bases from restriction enzyme cut sites as too close to the end of the DNA fragment (the suspect range). Invitrosomes that map within suspect range regions are considered theoretically subject to fragment end bias and so are defined as "suspect" nucleosomes. Invitrosomes that do not fall within the suspect range regions are assumed to not be affected by fragment end bias and are classified as "passed" nucleosomes. The restriction sites of the two restriction endonucleases used will usually not be near one another on the DNA. Therefore, invitrosomes from one experiment that are defined as suspect and normally would be discarded (due to proximity to a fragment end) can be recovered if the same locus is found to be occupied by a passed invitrosome in the second experiment. This is demonstrated in Figure 4.1 with the example invitrosomes in position 3 and position 4. In contrast, invitrosomes in position 2 remain in doubt as this position is near a fragment end in both experiments and both invitrosomes are suspect. Additionally, the positions where DNA fragments where generated that were too small to participate in reconstitution can be recovered. As the likelihood of this happening with both endonucleases digestions is small; a position lost in one experiment can be recovered if in the second experiment the fragment is of sufficient size to form a passed invitrosome (e.g. Figure 4.1 position 1).



Nucleosome Positions Confirmed (✓) in Doubt (?) or Lost (X)

Position	1	2	3	4
Hinc II only	✓	?	?	✓
Rsa I only	X	?	✓	?
Recovered	✓	?	✓	✓

Figure 4.1: Recovery of lost or suspect invitrosome (grey ovals) positions by use of separate samples cut by different restriction enzymes. R (Rsa I) and H (Hinc II) indicate restriction cut sites within the loci shown (top black bar)

In order to validate our approach, we have applied our recovery methods to the invitrosome datasets generated using the Caenorhabditis elegans genome described in Locke et al.[14]. Using the same strict suspect range of 200 bp from DNA fragment ends used in the Locke analysis, we find that we can recover the vast majority of discarded, suspect invitrosome positions. As the suspect range is decreased, the recovery rate increases proportionately. The percent recovery is also dependent on the number of total invitrosome position generated in both datasets. These results show that our method is capable of preventing the massive loss of data current nucleosome reconstitution studies are limited by.

4.3 Materials and Methods

Creation of Invitrosomes

The following is modified from [10]. Naked genomic DNA from wild-type *C. elegans* (N2 strain) was isolated by digesting flash-frozen worms with proteinase K (Roche, 2mg/ml final concentration) in worm lysis buffer (0.1M Tris HCl at pH 8.5, 0.1 M NaCl, 50 mM EDTA, 1% SDS) at 65°C for 45 min followed by phenol, phenol/chloroform, chloroform extraction and ethanol precipitation. RNA was removed with RNAse A (Roche) followed by phenol/chloroform, chloroform extraction and ethanol precipitation. DNA templates for both the Rsa I and Hinc II experiments, high-molecular weight genomic DNA was digested with 200 units of either restriction enzyme Rsa I or Hinc II (New England BioLabs) with the supplied buffers and 1X BSA (New England BioLabs). Digestion were carried out at 37°C for two hours followed by phenol, phenol/chloroform, chloroform extraction and ethanol precipitation. The complete digestions run on a 1% UltraPure Agarose (Invitrogen) gel. A continuous smear of fragments was seen for both digestions with a distribution of fragments lengths visually estimated to be centered upon and enriched around ∼850 bp and ∼3500 bp for the Rsa I and Hinc II digestions, respectively. The Rsa I and Hinc II DNA digestions were assembled with recombinant Xenopus histones into nucleosomes in a 1.1:1 molar ratio of DNA to histone octamer, such that on average one nucleosome bound to one molecule of

DNA.

Isolation of invitrosome core DNA fragments

The following is modified from [14, 33]. Invitrosome core DNAs from both Rsa I and Hinc II reconstitutions were isolated by diluting the respective invitrosomes into a buffer containing 5 mM MgCl2, 5 mM $CaCl_2$, 70 mM KCl and 10 mM Hepes at pH 7.9 and digesting with 20 units of MNase (Roche) resuspended at 1 units per μL for 15 min at room temperature. The digestion was stopped by adding an equal volume of 3% SDS, 100 mM EDTA and 50 mM Tris. Histones were removed by treating with one-tenth volume proteinase K (20 mg/ml in TE at pH 7.4) for 30 min at 50°C followed by phenol/chloroform and chloroform extractions and ethanol precipitation. Invitrosome DNA cores were assayed for complete digestion and isolated on a 2% UltraPure Agarose (Invitrogen) gel run at 100 V for one hour, followed by DNA extraction from the gel using a QIAquick Gel Extraction Kit (Qiagen).

Illumina library preparation and sequencing

The following is modified from [14, 33]. The Rsa I and Hinc II libraries were prepared by processing the invitrosome core DNA fragments using Illumina Genomic DNA Sample Prep Kit (Illumina 2007 Rev. A). Fragment end repair, adapter ligation and library amplification were all done according to the kit instructions with the following exceptions. Since our libraries are composed of ∼147 bp DNA cores rather than intact genomic DNA, the protocol was started at the "Perform End Repair" step. At this step, separate aliquots of both the Rsa I invitrosome core DNA sample and the Hinc II invitrosome core DNA sample were used. At the "Ligate Adapters to DNA Fragments" step, the purification was performed with the QIAquick PCR Purification Kit (Qiagen) rather than MinElute PCR Purification Kit. Additionally, a no-DNA control sample was processed in parallel to the Rsa I and Hinc II samples. After library preparation, each library was sequenced using a single lane of the Illumina GAII sequencer, resulting in 9.5 million and 5.5 million raw 36 bp reads for the Rsa

I and Hinc II libraries, respectively.

Mapping and preprocessing of reads

Once the raw reads were obtained, they were mapped back to the WS190 *C. elegans* genome using BLAT available on the BYU supercomputer. Parameters were set so that reads only mapped with sequences that would generate very high alignment scores (maximum of 1 mismatch). All other parameters were set to default. A number of reads were eliminated at this point because they failed to map back to provided template. The programs generated for use with this approach assume one position per read. Failure to do so introduces errors. To prevent this multiple alignments were processed using perl scripts that only allow match with the greatest alignment score to be retained.

Defining Suspect regions

Our recovery approach is composed of three steps the first of which is the generation of a suspect range based on a user-defined variable. To generating a suspect range, the exact position of the fragment start and end positions is required. We were able to define fragment start and ends by use the fragment end tables generated by Locke et al. and avaliable at http://nucleosome.rutgers.edu/nucenergen/celegansnuc/xfer [14]. These tables contain the start, end and fragment size of all hypothetical fragments generated across all chromosomes by both restriction endonucleases. However, the palindromic cuts sites are not included in the provided end/start positions. The size of the suspect range is limited by the number of bases from the fragment end that should be consider to be subject to fragment end bias. In the Locke analysis this was defined to by 200 bp from a fragment end and 200 bp from the fragment start, for a total range of 400 bps. For the purpose of assessing our approach, we defined multiple suspect ranges beginning at a minimum suspect range of a single bp and then increasing in specific increments until a max suspect range of 200 bp was reached. We chose a maximum suspect range of 200 bp so as to match the results of the Locke analysis.

To generate each suspect range, each fragment end and beginning position had the defined number of bases pairs added and subtracted from it. All positions that fell between the fragment start or end position and the end of the suspect range region were parsed into a unique output bedfile defined by the fragment set of origin (i.e Rsa I generated fragments or Hinc II generated fragments). The sets of positions found in these output files define all positions within the suspect range region and allowed us to separate them from the remainder of the fragment. The code for the perl script utilized at this step is provided in supplemental Figure A.15.

Defining Suspects and Passed nucleosomes

The second step in our approach is to define suspect and passed nucleosomes using the suspect ranges defined in the previous step. All the remaining reads post processing were compared to and defined by their location relative to the suspect range. The reads were classified as either suspect or cleared. If the read was found to begin within the suspect range, was classified as a suspect invitrosome. The sense of the read is taken into account when this comparison was made. All invitrosomes that do not receive this classification are considered cleared because they did not fall within the suspect. Once defined, the two classifications were separated into four temporary output files: Hinc II suspect invitrosomes, Hinc II cleared invitrosomes, Rsa I suspect invitrosomes, and Rsa I cleared invitrosomes. The code used in the step is provided in Figure A.14.

Recovery of suspect nucleosomes

The final step in our approach is recovery suspect invitrosomes by comparison to the alternative experiment's set of passed invitrosomes. For the purpose of our assessment, Rsa I suspect invitrosomes were compared to cleared Hinc II invitrosomes and Hinc II suspect invitrosomes were compared to cleared Rsa I invitrosomes. Suspect invitrosomes that sit at the same position as cleared invitrosomes in the alternative fragment set are now classified as recovered. Those that do not receive this new classification are considered to be biased invitrosomes. The final result is a set of recovered and biased invitrosomes for each fragment set. The code for step is provided in Figure A.14.

4.4 Results

4.4.1 Discarded Invitrosomes. The Locke et al. datasets we use in our analysis were derived from invitrosomes formed on *C. elegans* genomic DNA [14]. In their analysis, two separate invitrosome data set were made by reconstituting invitrosomes on *C. elegans* genomic DNA that had been digested with either Rsa I (a blunt, four-cutter) or with Hinc II (a blunt, five-cutter). Invitrosome core DNA was isolated using micrococcal nuclease and sequenced on the Illumina platform. The resulting Rsa I reconstitution experiment produced a total of 9.5 million raw sequencing reads, while the resulting Hinc II reconstitution experiment produced a total of 5.5 million raw sequencing reads. To control for invitrosomes positioned due to end effects, Locke et al. defined a 200 bp suspect range from each restriction enzyme cut site. In the *C. elegans* genome Rsa I cuts on average once per 490 bp, and Hinc II cuts on average once every 2109 bp. Use of their 200 bp suspect range resulted in excluding 87.7% of genomic bps for the Rsa I dataset and 19% of genomic bps for the Hinc II dataset, an alarmingly large portion of the genome [14].

We hypothesized that we could recover a significant portion of invitrosome positions lost to the Locke et al. analysis by applying our recovery approach. Thus we used the 9.5 million Rsa I raw sequencing reads and the 5.5 million Hinc II raw sequencing reads from Locke et al. in our analysis.

4.4.2 Pre-processing of reads. Because we were using raw reads, it was necessary to eliminate poor quality reads and reads that mapped to multiple loci before our approach could be applied. The raw reads were mapped back to the WS190 *C. elegans* genome using BLAT. Parameters were set so that reads only mapped with sequences that would generate very high alignment scores (maximum of one mismatch). A number of reads from both sets mapped to multiple sites within the genome. As our approach assumes one position per read, these multiple alignments were processed so only the match with the greatest alignment score was retained. Using these parameters, a total of 8.4 million (88.4%) of the original Rsa I generated sequence reads mapped back to the genome, while a total of 4.8 million (87.2%) of the original Hinc II sequence reads mapped to the genome.

4.4.3 Application of Approach. Our recovery approach is composed of three steps, namely

(i) a suspect range is generated based on a user-defined variable,

(ii) invitrosomes are mapped and declared either passed or suspect, and

(iii) suspect invitrosomes are recovered by comparison to the alternative experiment's set of passed invitrosomes.

In application of the first step, generation of suspect range regions is dependent on knowing precise fragment ends produced by restriction enzyme digestion. Because two different restriction endonucleases are used, the loci that fall into the suspect range regions will be different for the two experiments and will depend on the restriction enzyme use to prepare the template DNA for reconstitution. In applying this step in our analysis, we used the fragment end list generated by Locke et al. to define the beginning and end of DNA fragments based on the presence of either a Rsa I or a Hinc II cut site. This list contains the start, end and fragment size of all hypothetical fragments generated across all chromosomes by digestion with these enzymes. In the Locke analysis the suspect range was defined as 200 bp from a fragment start and 200 bp from the fragment end, a total range of 400 bps per DNA fragment [14]. For the purpose of assessing the efficacy of our approach, we defined multiple suspect ranges beginning at a minimum suspect range of a single bp and then increasing in defined increments until a maximum suspect range of 200 bp was reached. We chose the 200 bp maximum suspect range to match the results of the Locke analysis. To generate each suspect range region, the genomic position of each DNA fragment start or DNA fragment end (excluding the palindromic restriction enzyme cut site) had the suspect range-defined number of base pairs added to or subtracted from it respectively producing suspect range defined starts or ends. Any genomic regions between the original start of a DNA fragment and the suspect range defined start of that fragment, or the original end of a DNA fragment and the suspect range defined end of the DNA fragment were defined as the suspect range regions. This resulted in unique sets of suspect range regions for each restriction enzyme at each of the suspect range sizes.

We applied the second step of our approach by first mapping all the invitrosome sequence reads from both experiments to the WS190 version of the *C. elegans* genome. After mapping the sequence reads, each read was extended out to 147 bp to represent the entire footprint of the invitrosome from which it was derived and the start and end of each invitrosome was noted to produce Rsa I invtrosomes with both starts and ends and Hinc II invitrosomes with both starts and ends. The invitrosome starts and ends from both sets of reconstitutions were then compared to their respective suspect range regions. Depending on where each invitrosome falls relative to the suspect range regions (within the suspect range or outside of the suspect range), it is defined as either “suspect” or “passed”. Any invitrosome with a start that fell into suspect range start region or any invitrosome with an end that fell into a suspect range end region were defined as suspect. Passed invitrosomes were separated from suspect invitrosomes and kept as good data for each experiment. For each experiment the suspect range was kept the same between the Rsa I and the Hinc II datasets. This resulted in four invitrosome sets from the two experiments: passed Rsa I invitrosomes, suspect Rsa I invitrosomes, passed Hinc II invitrosomes and suspect Hinc II invitrosomes. The distribution of invitrosomes within those files are shown in Figures 4.2 and 4.3 below.

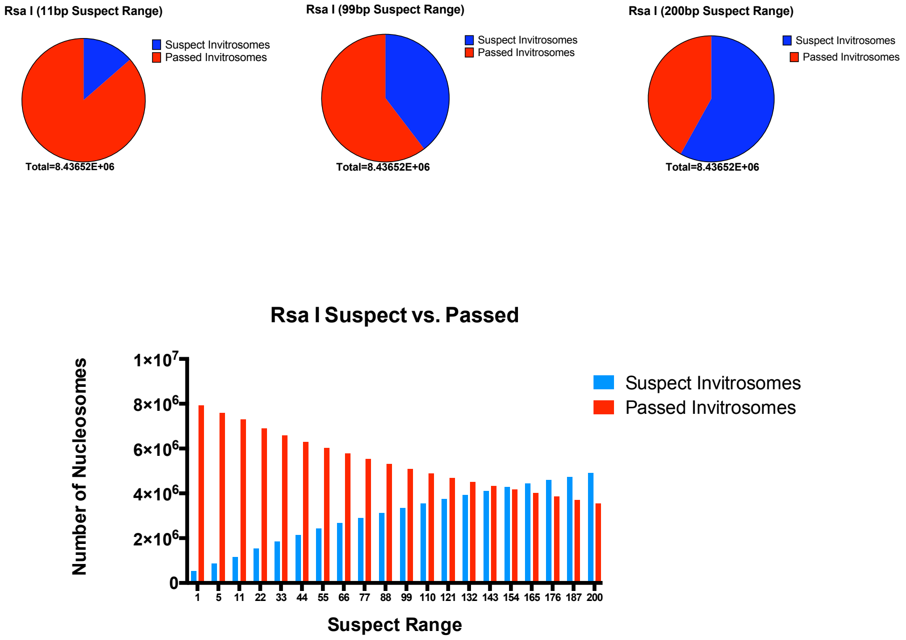


Figure 4.2: Number of valid invitrosomes decrease as the window size increases for RSAI data set. The end window is defined as the range of bases before or after restriction endonuclease sites that are considered susceptible to end bias. Nucleosomes that fall within said window are considered suspect and are eliminated from the data set. Nucleosomes that do not fall within this window are considered passed.

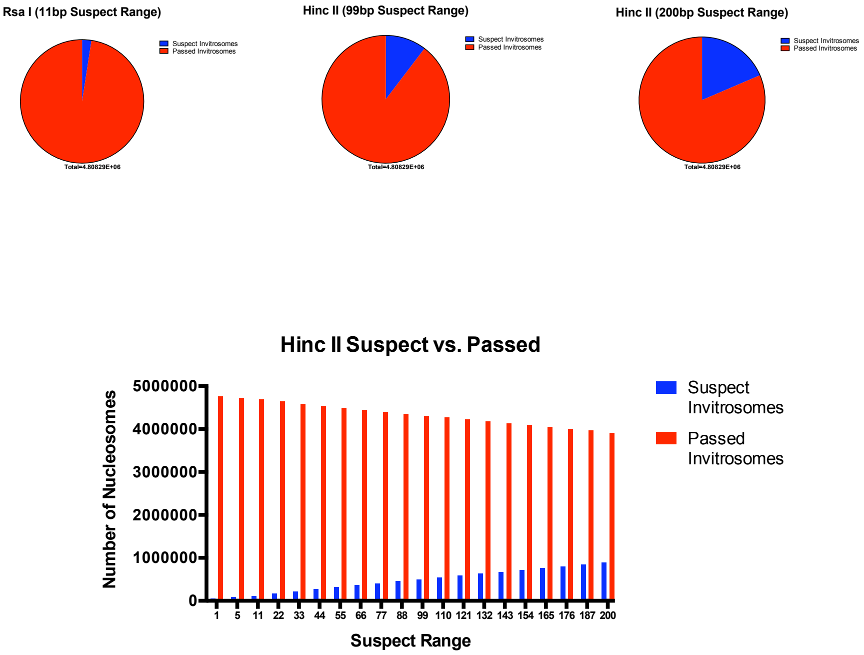


Figure 4.3: Number of passed nucleosomes decrease as the window size increases for Hinc II data set. Nucleosomes that fall within said window are considered suspect and are eliminated from the data set. Nucleosomes that do not fall within this window are considered passed.

Our final step was to recover suspect invitrosomes from one experiment that could be supported as being free of end-effect bias by comparison with passed invitrosomes reads from the alternative experiment. For the purpose of our assessment, Rsa I suspect invitrosomes were compared to passed Hinc II invitrosomes and Hinc II suspect invitrosomes were compared to passed Rsa I invitrosomes. Suspect invitrosomes that sit at the same position as passed invitrosomes in the alternative fragment set were now reclassified as "recovered" invitrosomes. Those that did not receive this new classification are considered to be potentially affected by end-bias and were reclassified as "biased" invitrosomes. The final result is a set of recovered and biased invitrosomes for each fragment set. The results generated by step two and this step were six unique output files: passed Rsa I invitrosomes, recovered Rsa I invitrosomes, biased Rsa I invitrosomes, passed Hinc II invitrosomes, recovered Hinc II invitrosomes, and biased Hinc II invitrosomes.

4.4.4 Recovery of Rsa I and Hinc II Invitrosomes. The mapped Rsa I dataset contained a total of 8,436,517 invitrosomes. Using our maximum suspect range of 200 bp; 4,899,372 or 58.1% of the Rsa I invitrosomes were declared suspect. Without our recovery method these suspect invitrosomes would be lost to further analysis. This is substantially lower than the number excluded from the Locke et al. analysis, but still a very large portion of the data.

In order to recover suspect Rsa I invitrosomes we compared these invitrosomes to the passed Hinc II invitrosomes that were analyzed at the Hinc II 200 bp-suspect range. As described above, any Rsa I suspect invitrosome that shared the same position with a Hinc II passed invitrosome was assumed to be an invitrosome that formed at that particular locus due to preferable DNA sequence rather than end-position bias and was declared recovered. This comparison resulted in 3,170,754 of the suspect Rsa I invitrosomes being reclassified as recovered. Thus using our recovery method we recovered 64.7% of the suspect Rsa I invitrosomes resulting in a total of 6,707,899 passed or recovered Rsa I invitrosomes, or

79.5% of the original invitrosome set. This left 1,728,618 suspect invitrosomes that were reclassified as biased and unusable, 20.5% of the original Rsa I invitrosome set, instead of the 58.1% that would be unusable without our recovery procedure. These results are shown below in Figure 4.4

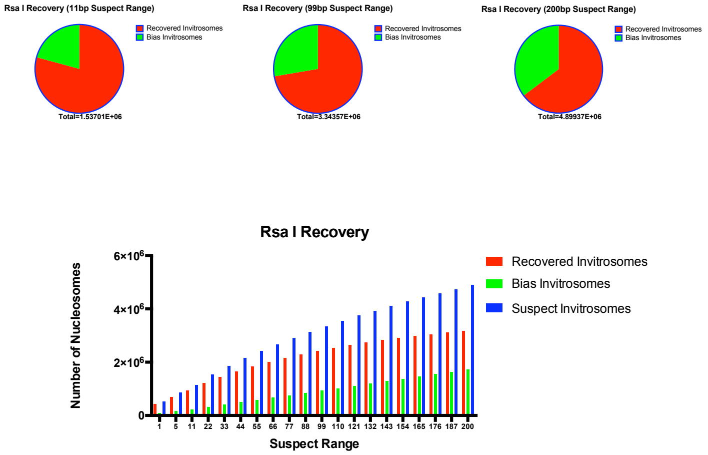


Figure 4.4: Suspect invitrosomes in the Rsa I dataset are recovered using comparative approach. The suspect invitrosomes (blue) within the Rsa I dataset are compared to passed invitrosomes within the Hinc II data set. This results in the recovery of 64.7% of the suspect invitronsomes, or recovered invitrosomes (red) while 34.3% remain and are considered bias (green).

The same analysis was performed on the 4,808,294 mapped Hinc II invitrosomes, with recovery analysis being performed with the passed Rsa I invitrosomes that were analyzed at the Rsa I 200 bp-suspect range. At a suspect range of 200 bp; 892,645 or 18.6% of the Hinc II invitrosomes were declared suspect. Using the passed Rsa I invitrosomes, 300,080 Hinc II suspect invitrosomes were recovered while the remaining 592,565 (66.4%) Hinc II suspect invitrosomes were labeled as biased. Thus using our recovery method we recouped 33.6% of the suspect Hinc II invitrosomes resulting in a total of 4,215,729 passed or recovered Hinc II invitrosomes, or 87.7% of the original invitrosome set. The remaining 592,565 biased invitrosomes represent 12.3% of the original Hinc II invitrosome set that was still unusable. Despite the more modest size of this recovery, it still represents a substantial improvement over the 18.6% that would be unusable without our recovery procedure. These results are shown below in Figure 4.5.

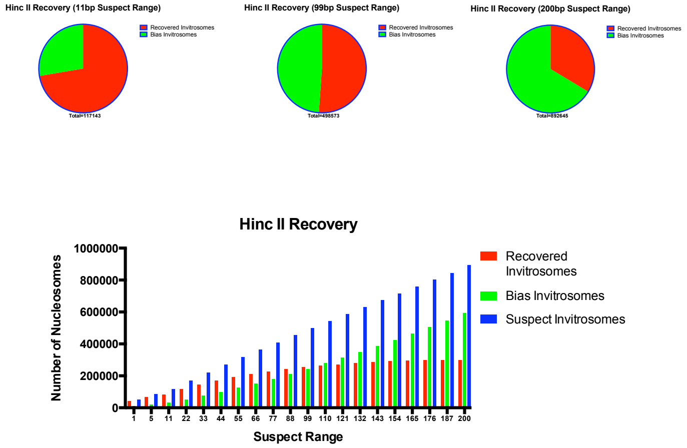


Figure 4.5: Suspect invitrosomes in the Hinc II dataset are recovered using comparative approach. The invalid invitrosomes (blue) within the Hinc II dataset are compared to passed invitrosomes within the Rsa I data set. This results in the recovery of 33.6% of the invalid invitrosomes (red) while 36.4% remain and are considered bias (green).

4.4.5 Varying the suspect range length.

We wanted to test the effect of varying lengths of suspect ranges on the number of invitrosomes declared suspect and recovered by our approach. Toward this end, we applied 19 more suspect ranges beginning with 1 bp, 5 bp, 11 bp (one helical turn of DNA) and then increasing by 11 bp until reaching 187 bp. We compared the results of applying these additional 19 suspect ranges to the results from our maximum 200 bp suspect range. As expected, we see that the number of suspect invitrosomes decreases linearly in relation to the length of the suspect range, and the lowest suspect range of a single base pair results in a low of only 515,035 (6.1%) of the Rsa I and 51,618 (1.1%) of the Hinc II invitrosomes being declared suspect respectively. It is interesting to note that at the larger suspect ranges (154 bp), for Rsa I invitrosomes, we observed that the number of suspect invitrosomes is actually greater than the passed invitrosomes. This is not the case for the Hinc II invitrosomes. All of these trends are demonstrated in Figures 4.4 and 4.5

One suspect range is of particular interest. The 11 bp suspect range represents one full turn of the DNA helix. If invitrosomes were to be affected by end-bias, but still try and retain a preferential rotational setting, it might be predicted that they would form between 1-11 bp from the end of the DNA fragment. Interestingly, it has been demonstrated that virtually all end-effect remodeling results in invitrosomes within about 10 bp of the fragment end [37] . At the 11 bp suspect range 1,145,346 (13.6%) of Rsa I invitrosomes are suspect and 117,143 (2.4%) of Hinc II invitrosomes are suspect. At this same level 922,900 (80.6%) and 84,717 (72.3%) of the suspect Rsa I and suspect Hinc II invitrosomes are recovered respectively. Having applied our approach, we find that a substantial number of suspect invitrosomes can be recovered within the Rsa I invitrosome set no matter what size the suspect range is. Within the maximum suspect range we find that our approach is able to recover 64.7% or 3,170,754 of the suspect invitrosomes within that particular suspect range. However, within the smaller range such as 11 bp we are able to recover 80.6% or 922,900 of the suspect invitrosomes within this suspect range. It should be noted that in the Locke analysis an 11 bp allowance was used when mapping the invitrosomes back to the genome. In all our previously described analyses we have used this same allowance when recovering suspect invitrosomes. That is to say, we reclassified a suspect invitrosome as recovered if the footprint of the suspect invitrosome overlapped with the footprint of a passed invitrosome from the alternative invitrosome set, effectively mapping within 11 bp (one helical turn). When this allowance is removed and an exact overlap is required, all previous described trends remain the same. The only observable difference is that actual recovery rates decrease by 3.2 - 3.5% for the Rsa I analyses and 1.2% to 2.9% for the Hinc II analyses across all suspect ranges, with the exception of the 1 bp suspect ranges where the decrease is 4.2% and 9.0% for Rsa I and Hinc II respectively.

4.5 Discussion

Our findings can be summarized in the statement of a few observed trends. First, recovery is most efficient when the suspect range is minimized. However, when a conservative suspect range is set, recovery is still significant. Of the two datasets, Rsa I achieves the greatest amount of recovery, but this is expected as it contained the larger number of suspect nucleosomes to begin with. The Hinc II, in contrast, had a much lower recovery rate, but also far fewer suspect invitrosomes. When the stringency of recovery was increases, that is to say when a perfect alignment was required for a suspect invitrosome to become a passed invitrosome, all observed trends in recovery rate and suspect nucleosome definition for both fragment sets remains the same.

Of note is the plateau in recovery rate observed in the Hinc II dataset (Figure 4.5). We believe that this due to the difference in the number of invitrosome avaliable for recovery between the datasets. The Hinc II data is smaller than the Rsa I dataset. Additionally, inherent to our approach is that as the suspect range increases, the number of passed nucleosomes avaliable for the recovery of the other set decreases. We have found that in both the Rsa I and Hinc II dataset, a liner relationship exists between the recovered invitrosomes and the suspect ranged used when the passed invitrosome set used for for recovery is factored in. We believe the observed plateau to be a reflection of this linear relationship and the decreasing passed Rsa I invitrosome set.

Thus present a novel method by which end-bias can be successfully addressed while eliminating the need to discard large portions of data. This method generates two separate sets of fragments using two different restriction enzymes, which are used in standard nucleosome reconstitutions. Suspect invitrosomes that would normally be discarded can be recovered by comparing the two dataset using our three-step approach. First a suspect range is defined and all positions are mapped to the sequences from which they originated. Invitrosomes are then separated based on where they fall relative to the suspect range. Finally, suspect invitrosomes from one invitrosome set that would usually be discarded are compared to the passed invitrosomes from the second fragment set. If a passed invitrosomes is found at the same position as the suspect invitrosome in the first set, the suspect nucleosome is determined to have formed there by a force other than fragment end bias and is included in the usable dataset. By applying our method to the invitrosome data sets generated by Locke et al., we have demonstrate that this is a valid approach for substantial data recovery and thus provide a more complete dataset for analysis. With a more complete dataset, studies utilizing reconstituted nucleosomes will be able to provide a more complete insight into the influence of intrinsic sequence on the positions of nucleosomes

4.6 Future Directions

We have prepared this work for publication and a complete manuscript has been written. This work is the first of its kind and represent a unique method of recovering *in vitro* nucleosome reconstitution data. The data utilized in this study was the first to genome-wide invitrosome generate datasets with clearly defined fragment ends. While having defined ends does introduce known bias, this same information allows for very significant recovery of these types of datasets. We plan to submit this publication to PloS one and BMC Genomics July 1st, 2014 and believe it will be a new standard for genome-wide invitrosome analysis.

Bibliography

[1] J.C Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[2] Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[3] G. Brumfiel. The astronomical unit gets fixed. *Nature News*, 2012.

[4] B. Alberts et al. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[5] R. D. Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–71, 1974.

[6] K. Struhl and E. Segal. Determinants of nucleosome positioning. *Nat Struct Mol Biol*, 20(3):267–73, 2013.

[7] A. Annunziato. Dna packaging: Nucleosome and chromatin. *Nature Eduation*, 1(1):1, 2008.

[8] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–60, 1997.

[9] J. Bendar et al. Nucleosomes, linker dna, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences*, 95(24):14173–14178, 1998.

[10] S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, and A. Z. Fire. Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. *Genome Res*, 16(12):1505–16, 2006.

[11] A. Miller. Human gene therapy comes of age. *Nature*, 357(6378):455–460, 1992.

[12] S. M. Johnson. Painting a perspective on the landscape of nucleosome positioning. *J Biomol Struct Dyn*, 27(6):795–802, 2010.

[13] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, 2006.

[14] G. Locke, D. Haberman, S. M. Johnson, and A. V. Morozov. Global remodeling of nucleosome positions in c. elegans. *BMC Genomics*, 14:284, 2013.

[15] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–20, 2011.

[16] J. Y. Lee and T. H. Lee. Effects of dna methylation on the structure of nucleosomes. *J Am Chem Soc*, 134(1):173–5, 2012.

[17] S. Pennings, J. Allan, and C. S. Davey. Dna methylation, nucleosome formation and positioning. *Brief Funct Genomic Proteomic*, 3(4):351–61, 2005.

[18] C.K Collings, P.K Waddell, and Anderson J.N. Effects of dna methylation on nucleosome stability. *Nucleic Acids Res*, 41(5):2918–31, 2013.

[19] P. Chen et al. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev*, 27(19):2109–24, 2013.

[20] B. Guillemette et al. Variant histone h2a.z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol*, 3(12):e384, 2005.

[21] C. Jin and G. Felsenfeld. Nucleosome stability mediated by histone variants h3.3 and h2a.z. *Genes Dev*, 21(12):1519–29, 2007.

[22] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.

[23] A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res*, 20(1):90–100, 2010.

[24] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*, 18(7):1073–83, 2008.

[25] G. Keller. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev*, 19(10):1129–55, 2005.

[26] H. Niwa et al. Self-renewal of pluripotent embryonic stem cells is mediated via activation of stat3. *Genes and Dev.*, 12:2048–2060, 1998.

[27] H. Niwa et al. Quantitative expression of oct-3/4 defines differentiation, dedifferentiation or self-renewal of es cells. *Nat. Genet*, 24:372–376, 2000.

[28] K. Mitsui et al. The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and es cells. *Cell*, 113:631–642, 2003.

[29] A.G. Smith. Embryo-derived stem cells: Of mice and men. *Annu. Rev. Cell Dev. Biol.*, 17:435–462, 2001.

[30] R.H. Xu et al. Bmp4 initiates human embryonic stem cell differentiation to trophoblast. *Nat. Biotechnol*, 20:1261–1264, 2002.

[31] A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83, 2011.

[32] V. B. Teif, Y. Vainshtein, M. Caudron-Herger, J. P. Mallm, C. Marth, T. Hofer, and K. Rippe. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol*, 19(11):1185–92, 2012.

[33] G. Locke, D. Tolkunov, Z. Moqtaderi, K. Struhl, and A. V. Morozov. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proc Natl Acad Sci U S A*, 107(49):20998–1003, 2010.

[34] C.L. Smith and C.L Peterson. Atp-dependent chromatin remodeling. *Curr. Topic Dev. Bio.*, 65:115–148, 2005.

[35] Y. Xi et al. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome Res*, 21(5):718–24, 2011.

[36] P.N. Dyer et al. Reconstitution of nucleosome core particles from recombinant histones and dna. *Methods Enzymol*, 375:23–44, 2004.

[37] A. Flaus and T. J. Richmond. Positioning and stability of nucleosomes on mmtv 3'ltr sequences. *J Mol Biol*, 275(3):427–41, 1998.

[38] T. Sakaue, K. Yoshikawa, S. H. Yoshimura, and K. Takeyasu. Histone core slips along dna and prefers positioning at the chain end. *Phys Rev Lett*, 87(7):078105, 2001.

[39] P.N. Grokhovsky et al. Sequence-specific ultrasonic cleavage of dna. *Biophysical Journal*, 100:117–125, 2011.

[40] S.L Schwartz and M.L. Farman. Systematic overrepresentation of dna termini and underrepresentation of subterminal regions among sequencing templates prepared from hydrodynamically sheared linear dna molecules. *BMC Genomics*, 11:86, 2011.

Appendix A. Supplemental Figures

A.1 Supplemental Figures for Chapter 3

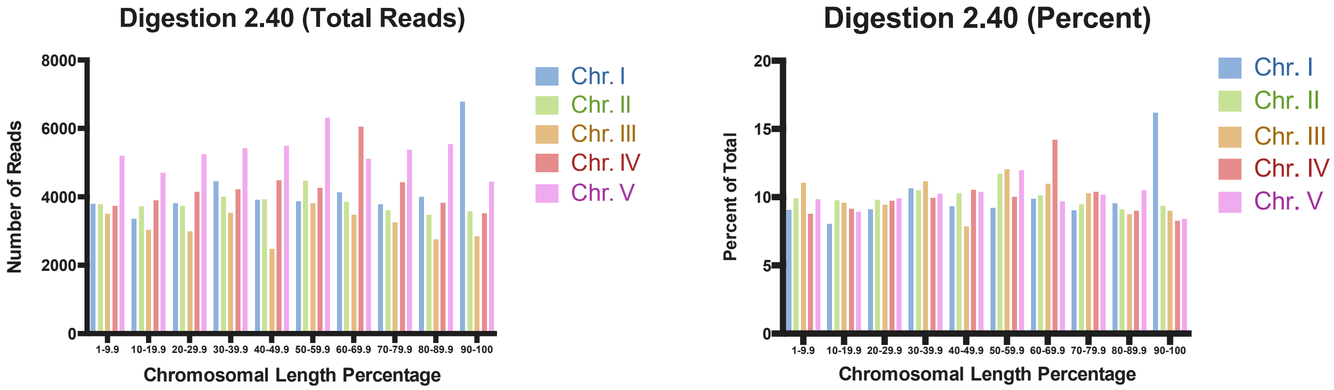


Figure A.1: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 3 of 3.3.

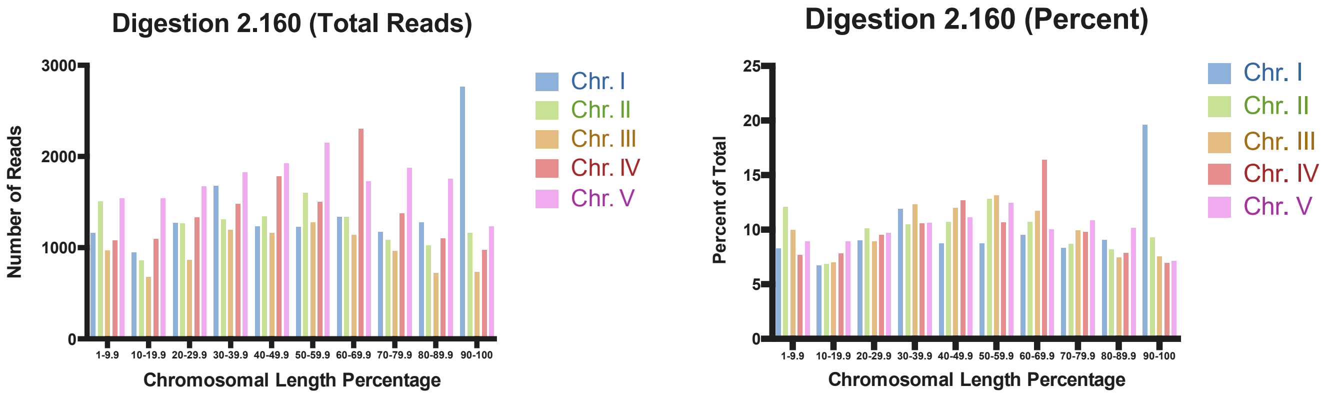


Figure A.2: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 4 of 3.3.

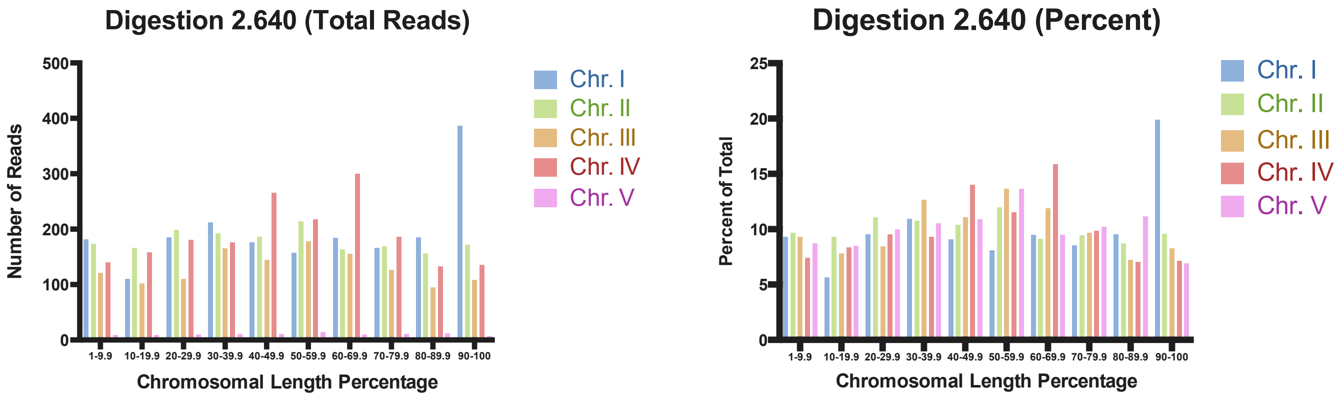


Figure A.3: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 5 of 3.3.

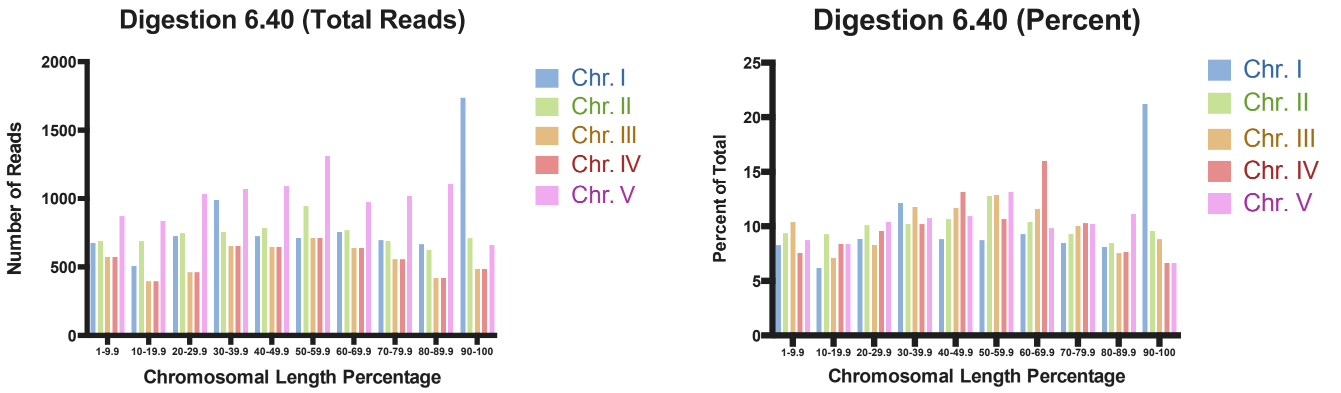


Figure A.4: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 7 of 3.3.

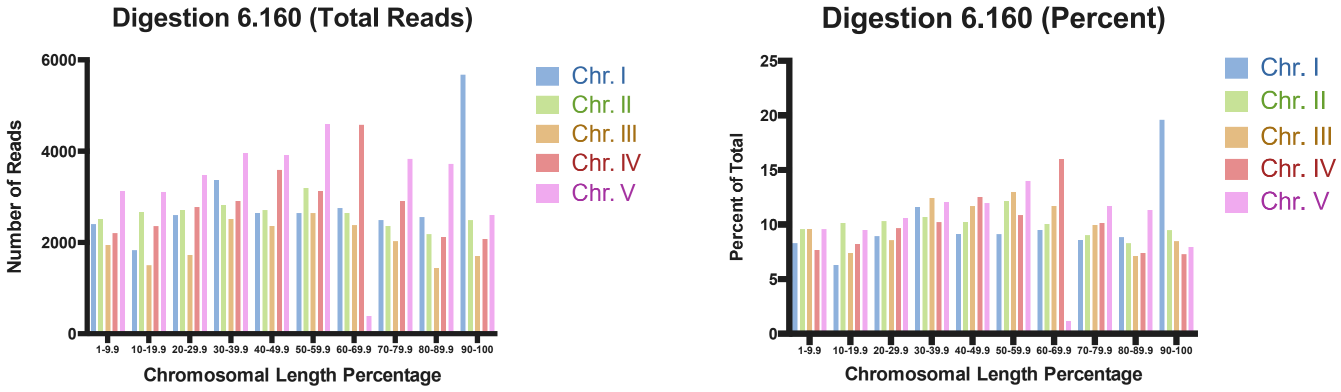


Figure A.5: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 8 of 3.3.

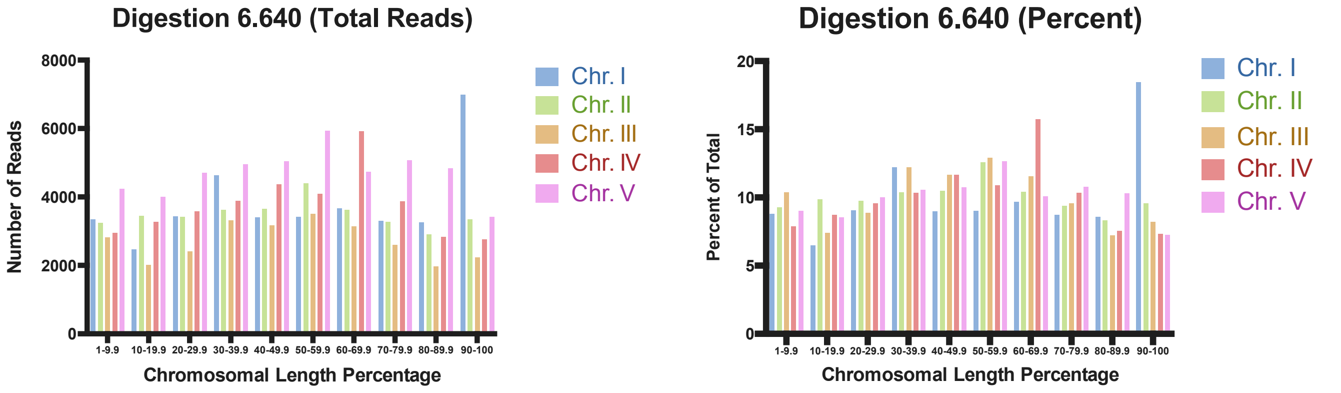


Figure A.6: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the heavy digestion corresponding to lane 9 of 3.3.

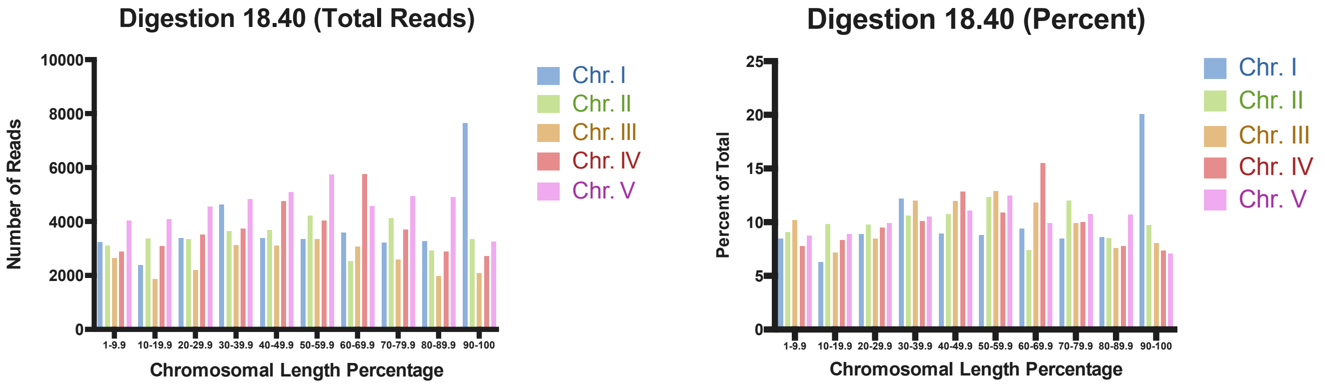


Figure A.7: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the light digestion corresponding to lane 11 of 3.3.

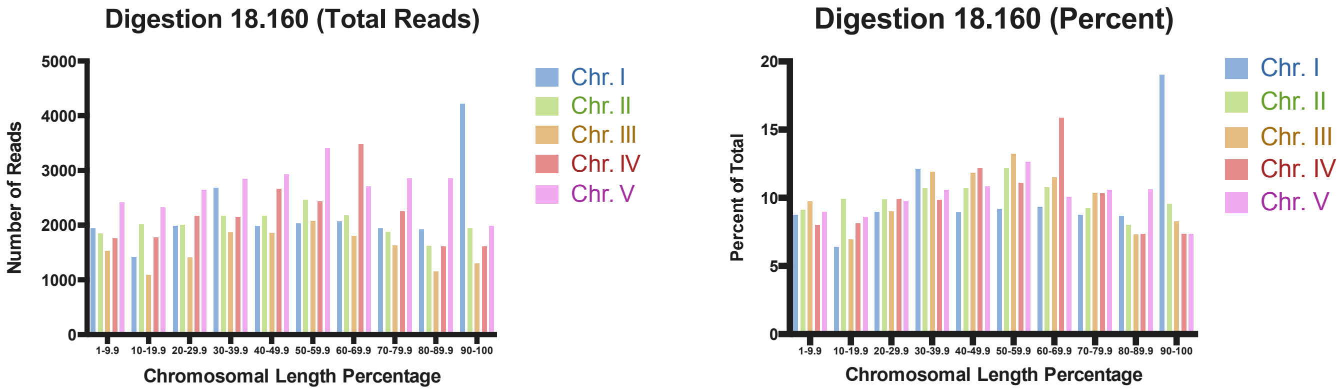


Figure A.8: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the heavy digestion corresponding to lane 12 of 3.3.

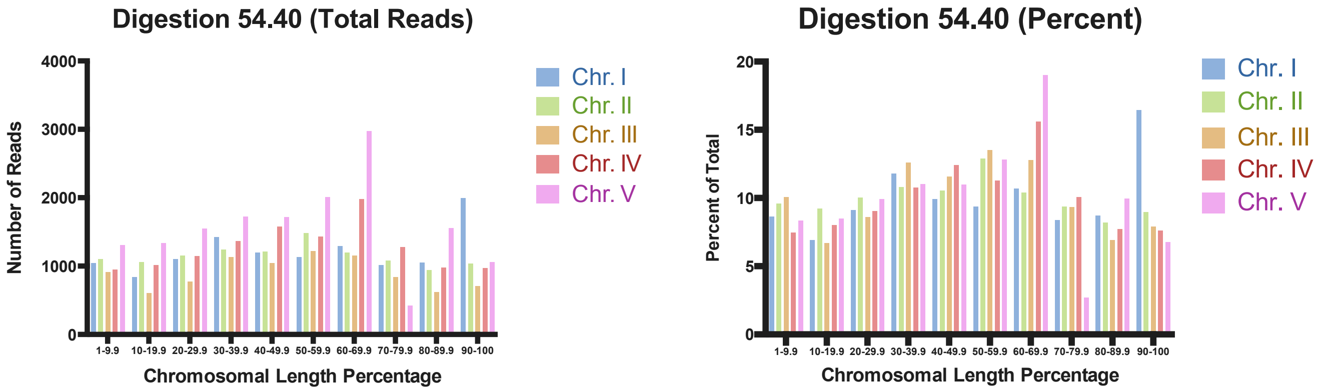


Figure A.9: Graphs show the total reads (left) and percent reads (right) that mapped to each tenth of the five *C. elegans* autosomes for the heavy digestion corresponding to lane 15 of 3.3.

A.2 Supplemental Figures for Chapter 4

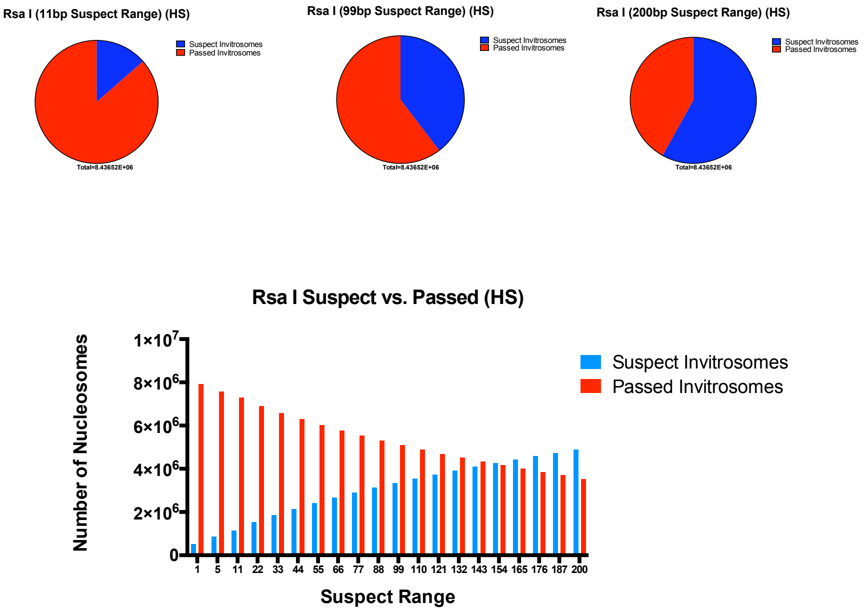


Figure A.10: Figure shows trends when a perfect alignment is required for definition. These trends match what is seen when a 11 bp allowance is used in analysis of Rsa I dataset.

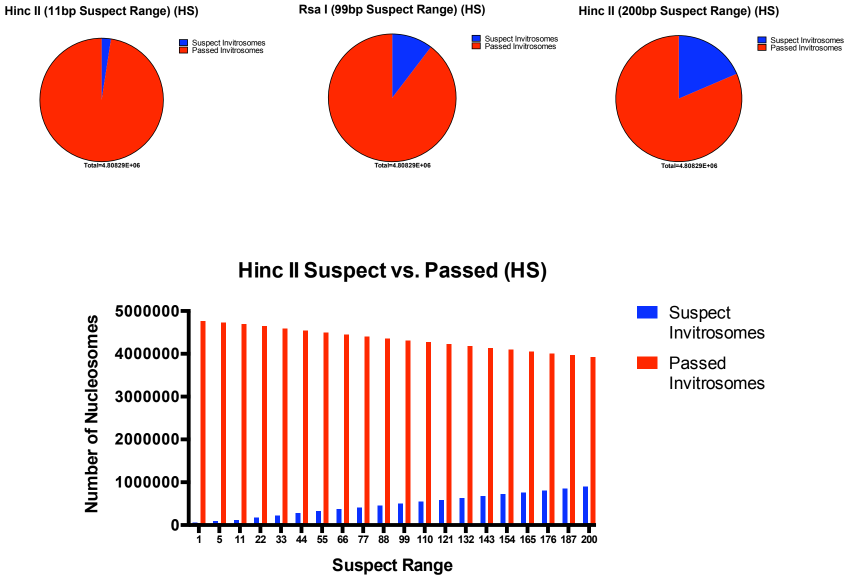


Figure A.11: Figure shows trends when a perfect alignment is required for definition. These trends match what is seen when a 11 bp allowance is used in analysis of Hinc II dataset.

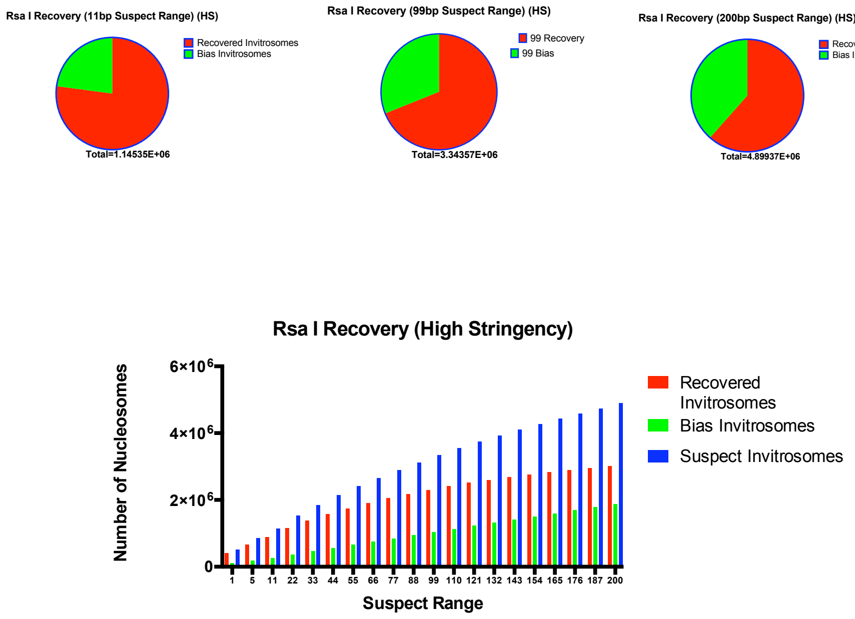


Figure A.12: Figure shows recovery trends when a perfect alignment is required for suspect recovery. These trends match what is seen when a 11 bp allowance is used in analysis of Rsa I dataset.

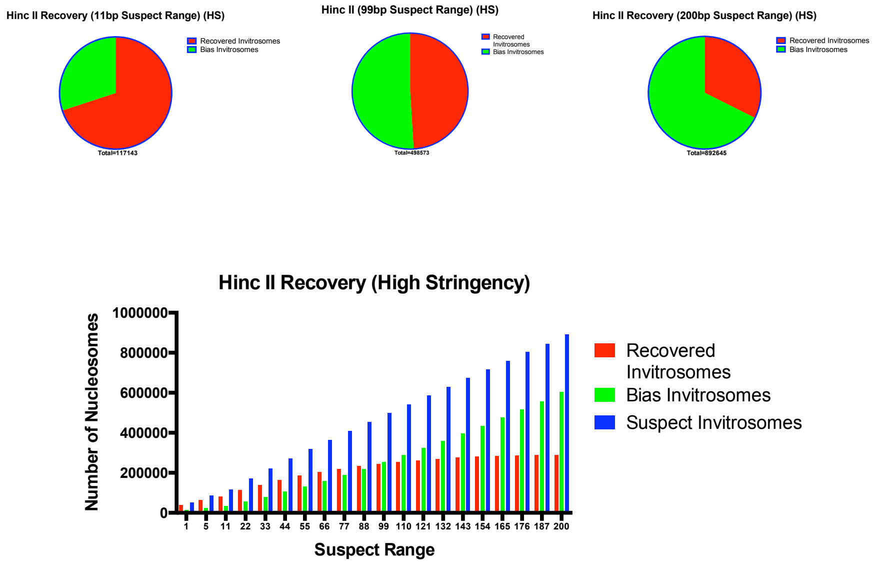
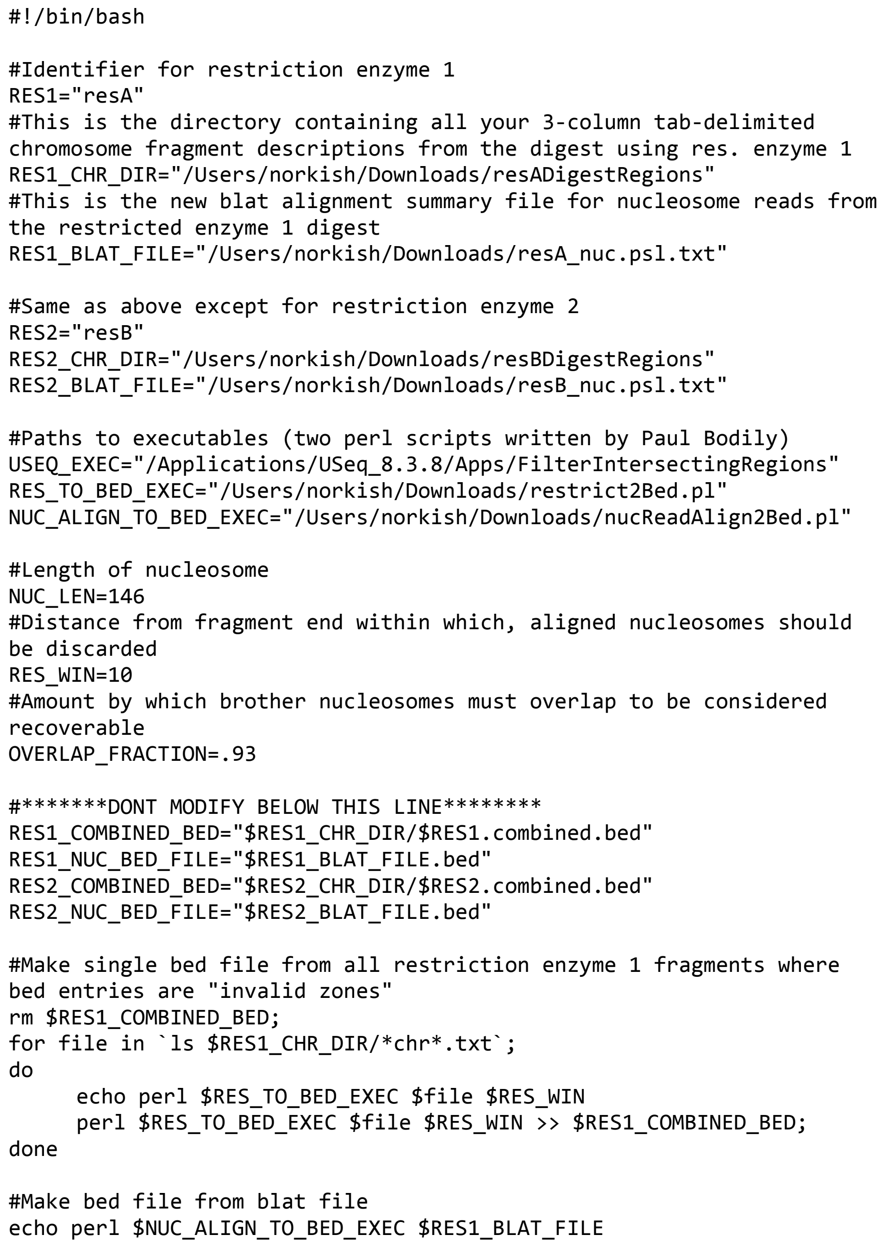
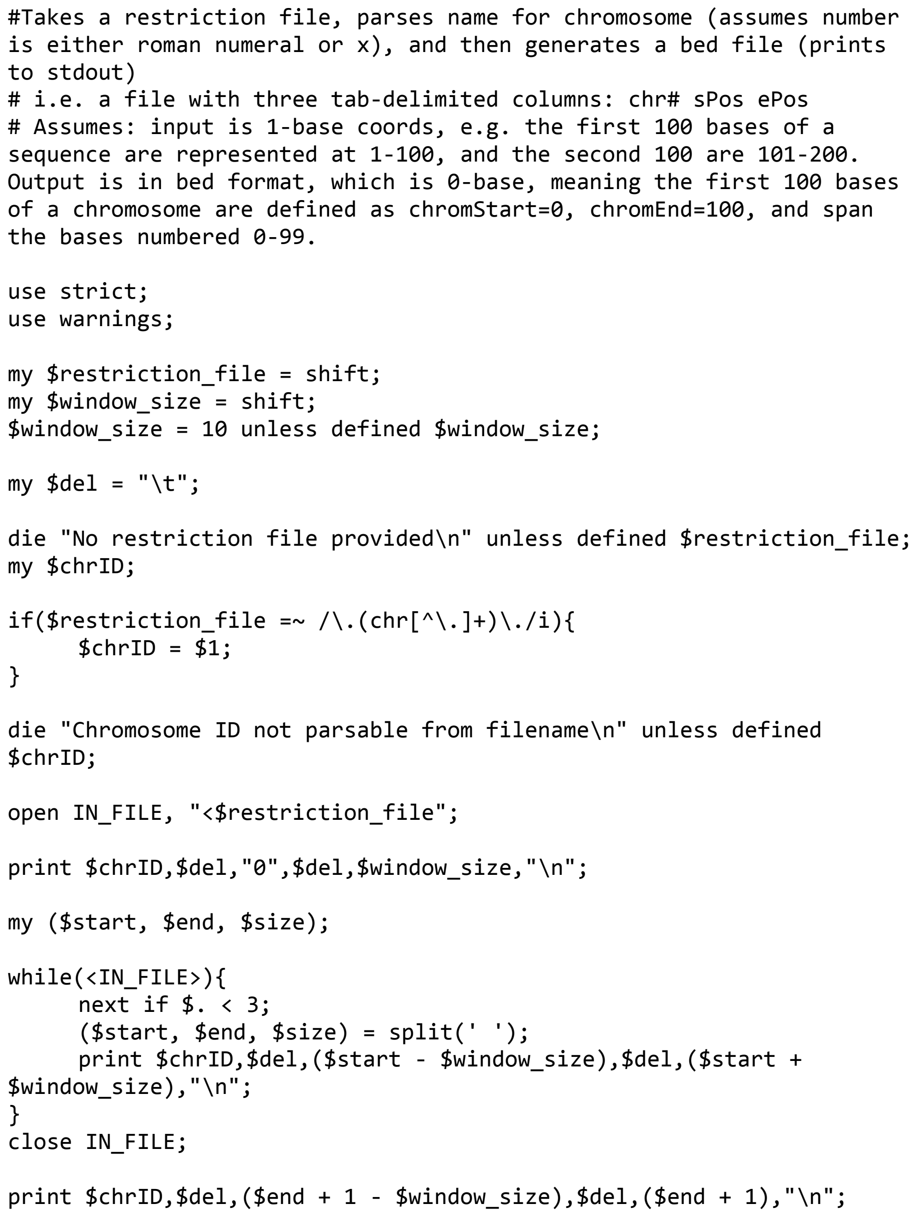


Figure A.13: Figure shows recovery trends when a perfect alignment is required for suspect recovery. These trends match what is seen when a 11 bp allowance is used in analysis of Hinc II dataset.



```
#!/bin/bash

#Identifier for restriction enzyme 1
RES1="resA"
#This is the directory containing all your 3-column tab-delimited
chromosome fragment descriptions from the digest using res. enzyme 1
RES1_CHR_DIR="/Users/norkish/Downloads/resADigestRegions"
#This is the new blat alignment summary file for nucleosome reads from
the restricted enzyme 1 digest
RES1_BLAT_FILE="/Users/norkish/Downloads/resA_nuc.psl.txt"

#Same as above except for restriction enzyme 2
RES2="resB"
RES2_CHR_DIR="/Users/norkish/Downloads/resBDigestRegions"
RES2_BLAT_FILE="/Users/norkish/Downloads/resB_nuc.psl.txt"

#Paths to executables (two perl scripts written by Paul Bodily)
USEQ_EXEC="/Applications/USeq_8.3.8/Apps/FilterIntersectingRegions"
RES_TO_BED_EXEC="/Users/norkish/Downloads/restrict2Bed.pl"
NUC_ALIGN_TO_BED_EXEC="/Users/norkish/Downloads/nucReadAlign2Bed.pl"

#Length of nucleosome
NUC_LEN=146
#Distance from fragment end within which, aligned nucleosomes should
be discarded
RES_WIN=10
#Amount by which brother nucleosomes must overlap to be considered
recoverable
OVERLAP_FRACTION=.93

#*******DONT MODIFY BELOW THIS LINE********
RES1_COMBINED_BED="$RES1_CHR_DIR/$RES1.combined.bed"
RES1_NUC_BED_FILE="$RES1_BLAT_FILE.bed"
RES2_COMBINED_BED="$RES2_CHR_DIR/$RES2.combined.bed"
RES2_NUC_BED_FILE="$RES2_BLAT_FILE.bed"

#Make single bed file from all restriction enzyme 1 fragments where
bed entries are "invalid zones"
rm $RES1_COMBINED_BED;
for file in `ls $RES1_CHR_DIR/*chr*.txt`;
do
      echo perl $RES_TO_BED_EXEC $file $RES_WIN
      perl $RES_TO_BED_EXEC $file $RES_WIN >> $RES1_COMBINED_BED;
done

#Make bed file from blat file
echo perl $NUC_ALIGN_TO_BED_EXEC $RES1_BLAT_FILE
```

Figure A.14: Code for perl program “Classifynucleosome.sh” written by Paul Bodily



```
#Takes a restriction file, parses name for chromosome (assumes number
is either roman numeral or x), and then generates a bed file (prints
to stdout)
# i.e. a file with three tab-delimited columns: chr# sPos ePos
# Assumes: input is 1-base coords, e.g. the first 100 bases of a
sequence are represented at 1-100, and the second 100 are 101-200.
Output is in bed format, which is 0-base, meaning the first 100 bases
of a chromosome are defined as chromStart=0, chromEnd=100, and span
the bases numbered 0-99.

use strict;
use warnings;

my $restriction_file = shift;
my $window_size = shift;
$window_size = 10 unless defined $window_size;

my $del = "\t";

die "No restriction file provided\n" unless defined $restriction_file;
my $chrID;

if($restriction_file =~ /\.(chr[^\.]+)\./i){
	$chrID = $1;
}

die "Chromosome ID not parsable from filename\n" unless defined
$chrID;

open IN_FILE, "<$restriction_file";

print $chrID,$del,"0",$del,$window_size,"\n";

my ($start, $end, $size);

while(<IN_FILE>){
	next if $. < 3;
	($start, $end, $size) = split(' ');
	print $chrID,$del,($start - $window_size),$del,($start +
$window_size),"\n";
}
close IN_FILE;

print $chrID,$del,($end + 1 - $window_size),$del,($end + 1),"\n";
```

Figure A.15: Code for perl program “restrict2bed.pl” written by Paul Bodily