Analysis of nucleosomal DNA patterns around transcription factor binding sites

Sheethal Umesh Nagalakshmi
su2484@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Thesis Research Project

Analysis of nucleosomal DNA patterns around transcription factor binding sites

Authored by Sheethal Umesh Nagalakshmi

Advised by Dr. Feng Cui

Committee Members: Dr. Gregory Babbitt, Dr. Gary Skuse, Dr. Feng Cui

This thesis is submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics

Department of Thomas H. Gosnell School of Life Sciences

College of Science

Rochester Institute of Technology

Rochester, NY

January 7th, 2021
Committee Members: Dr. Gregory Babbitt, Dr. Gary Skuse, Dr. Feng Cui

A. Abstract
B. Introduction
C. Materials and Methods
- Calculation of Nucleosome occupancy profile at and around the TF binding sites
- Calculation of ΔNPS at or around TF binding sites
- Data Overview
- Nucleosome positions data mapped by chemical method for mESCs and yeasts
- Table 1. Nucleosome positions data mapped by chemical method for mESCs and yeasts
- CHIP Seq datasets for mESCs from NCBI
- Table 2. CHIP Seq datasets for mESCs from NCBI
- Peak Calling using HOMER for TFs using GEO datasets in mESCs
- Table 3. Peak Calling using HOMER for TFs using GEO datasets in mESCs
- Peak calling using HOMER for TFs using Array Express datasets in yeasts
- Table 4. Peak calling using HOMER for TFs using Array Express datasets in yeasts
D. Results
- Table 5. Peak/dip conformations of TFs around ChIP fragment center in mESCs
- Table 6: Peak/dip conformations of TFs around ChIP fragment center in yeasts
- Table 8: Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in mESCs for validated pioneer factors (57)
- Schematic model of research findings
E. Discussion
F. Future work
G. Literature Cited
H. Supplemental Materials
- Nucleosome occupancy and ΔNPS profiles for TF binding sites in mESCs
- Nucleosome occupancy and ΔNPS profiles for TF binding sites in yeasts
A. Abstract

Nucleosomes are ~147bp DNA wrapped around the histone octamer which are involved in regulating gene transcription. They have the ability to disassemble depending on the process they are involved in and the nucleosome positioning controls the output of the genome. Therefore, it is important to understand the nucleosome positioning and how its positioning affects the binding of transcription factors (TFs) and gene expression thereby regulating the transcription outcome of the genome. Many studies suggest that TFs and nucleosomes compete with each other for genome accessibility. However, the majority of the studies focus on the nucleosome organization rather than underlying DNA sequences and its patterns which might actually be playing an important role in understanding the regulatory role of nucleosomes in gene transcription. This research study focuses on identifying the specific sequence patterns at or around TF binding sites. The study specifically focuses on identifying the fraction of nucleosomes with WW/SS and anti - WW/SS sequence patterns as they might be responsible for maintaining the stability of the nucleosomes. This will provide a new molecular mechanism underlying NDR formation around TF binding sites and pioneer TF-induced chromatin opening.
B. Introduction

The basic units of chromatin are nucleosomes that are capable of sliding along the DNA and regulating gene expression (1-3). Nucleosomes are ~147bp of DNA wrapped around the histone octamer that is made up of H2A, H2B, H3 and H4 core histones present in 2 copies each (3-6). Nucleosomes are involved in regulating cellular processes like DNA replication, DNA repair and gene transcription (1,7). They have the ability to completely or partially disassemble depending on the process they are involved in (1,7). Therefore, it is very important to understand nucleosome dynamics and the factors influencing them.

Studies have shown that nucleosome dynamics are influenced by nucleosome occupancy, histone modifications, and nucleosome positioning within the chromatin (1). “Nucleosome occupancy is the average number of nucleosomes” (1) present at a particular genome coordinate that influences the DNA’s availability to bind proteins at their target sites, thereby influencing chromatin function (1). Histone post translational modifications, like acetylation, methylation, phosphorylation, ADP ribosylation, glycosylation, sumoylation and, ubiquitinylation, influence DNA accessibility to transcription factors (TFs) (8-10). Nucleosome positioning refers to where the nucleosomes are selectively positioned and organized across the genome to regulate genomic function (11).

Previous studies in yeast have shown the presence of nucleosome depleted regions in the promoters that are transcriptionally active and enhancers which ensures the proteins their accessibility to DNA as well as regulating transcription and replication processes (11,12). In yeast, although the nucleosomes were found to be present in the promoter of stress regulating genes, their depletion was only seen during the gene activation, suggesting the role of nucleosome movement and their structural alterations in regulating the gene expression (1,13).
Nucleosome positioning controls the output of the genome (5). Therefore, it is important to understand nucleosome positioning and how its positioning affects the binding of transcription factors and gene expression thereby regulating the transcription outcome of the genome.

Nucleosomes present at a specific genomic coordinate can be measured using methods like CHIP-Seq, MNase and chemical methods (1). The most widely used methods to map nucleosome positions are micrococcal nuclease (MNase) and chemical methods (14). MNase is an endo-exo nuclease that digests the linker DNA between nucleosomes and continues digestion of the residual linker until there is loss of H1 resulting in the 147 bp nucleosomal DNA followed by high throughput sequencing (14, 15). In the chemical method, the histone H4 that contains a serine at position 47 in the histone octamer is first mutated to cysteine (H4S47C) (14). Next the cells are labelled with a sulfhydryl reactive copper chelating reagent followed by addition of the copper ions and hydrogen peroxide resulting in symmetrical cleavage of the DNA backbone leaving behind the nucleosome center (14,16 - 18). The problem with the MNase method is that the enzyme has a strong preference for A/T rich sequences that leads to sequence bias resulting in underrepresentation of nucleosomes with sequences that are A/T rich (14, 19). The chemical method is capable of finding the nucleosome center positions without sequence bias, but they would require the H4S47C mutation (14). However, this can be solved by replacing the histone H4 with H4S47C (14).

Previous studies have demonstrated that TFs compete with histone octamers for their target sites across the genome (20). Therefore, promoters in all the species are observed to lack nucleosomes and an anti-correlation is seen between their gene expression levels and nucleosome occupancy (20). However, chemical mapping has shown that not all TFs will compete to bind on the nucleosomal DNA as some might function as pioneer factors (14). This was demonstrated in the recent in vitro studies where certain TFs like Oct4, Sox2, Klf4 are
capable of binding directly to their target sites on the nucleosomes (14, 21). Pioneer factors are transcription factors that are capable of binding independently to their target sites present on the nucleosomial DNA even in closed chromatin condition where high nucleosome occupancy is observed (22). These regions cannot be accessed by other TFs as they do not have independent chromatin binding ability (22). It is necessary to learn more about pioneer factors as regulating their expression is important in various cancers (22, 23). Therefore, it becomes important to study nucleosomal DNA which might influence nucleosome positioning thereby influencing the pioneer TFs like Oct4, Sox2, Klf4 to directly bind to their targets.

Nucleosomal organization and their structural stability is greatly dependent on specific DNA patterns (24). Certain combinations of dinucleotides have the ability to favour or disfavour nucleosome formation by adjusting the bending of DNA either making them more bendable or resist them from bending (25-29). It is generally observed that nucleosomes are present in abundance in the GC rich regions whereas nucleosome depleted regions are observed in AT rich regions (30). Also, nucleosome positioning differs in unicellular and multicellular organisms. TFs can easily bind to their targets in unicellular organisms because of the presence of open genomic sites as long as the target sites are unrestricted by nucleosome repositioning (31). In multicellular organisms, the DNA sequences present will contribute to the stable nucleosome structures favoring nucleosomes particularly observed in promoters, whereas the instability of nucleosomes is observed in the unicellular organisms influenced by disfavoring sequences (31).

Nucleosome positions have preferences (32). They can have rotational positioning in which the DNA helix side faces towards the histone or transitional positioning which is the midpoint of the nucleosome with regard to the DNA sequence (32). Understanding this becomes important as accessibility to the DNA sequences on the nucleosomes and its activity is
determined by these 2 parameters (33). Also, these nucleosomes are observed to be placed side by side at a fixed distance from each other (33).

Previous studies in yeast have demonstrated that if the binding site of TF is considered nucleosome 0, the first nucleosome that is located upstream of the transcriptional start site (TSS) is considered as nucleosome -1 (33). This nucleosome will give access for binding of the promoter regulatory elements by undergoing changes that result in its destabilization, thereby helping in initiating the transcription (33). This nucleosome might be evicted for a long time until multiple rounds of transcription occur, or it might reassemble to its original structure between each transcription cycle (33). However, this question remains unanswered. The regions downstream of nucleosome -1 are the nucleosome flanking regions (33). The region downstream of nucleosome 0 is nucleosome +1 and that gets evicted while transcription takes place i.e., there is a nucleosome depletion region formation here and this will eventually return to its original location after RNA polymerase II passes (33). This is only observed in genes that are highly transcribed (34). Always, we see that the nucleosome -1, nucleosome 0 and nucleosome +1 are tightly positioned (33). The sequences might be favorable or unfavorable for nucleosome positioning (33).

Dinucleotides like AA, TT and GC are found to be periodically present and they provide a rotational setting of DNA wrapped around the histone octamer (33). AA and TT dinucleotides occurring in the major groove help in wrapping DNA around the histone core by expanding the major groove and the minor groove is contracted by GC dinucleotides (33,35,36).

The most widely described nucleosomal DNA sequence pattern is the WW/SS pattern (where W is A or T and S is G or C) (37, 41), where WW dinucleotides occur at sites where it bends into the minor groove and SS dinucleotides occur at DNA sites where it bends into the major groove (37, 38). Recently, a study showed the presence of an anti-WW/SS pattern in
promoter nucleosomes in yeast (39,40). It is demonstrated that yeast promoters follow the antiWW/SS pattern compared to conventional WW/SS pattern as these sequence patterns might contribute to unfavorable interaction between DNA and histone octamer, thus representing a relatively unstable structure (40,41).

Many studies suggest that TFs and nucleosomes compete with each other for genome accessibility (42 - 45). However, the majority of the studies focus on the nucleosome organization rather than underlying DNA sequences and its patterns which might actually be playing an important role in understanding the regulatory role of nucleosomes in gene transcription. Poly(dAdT) tracts are one of the major determinants of nucleosome organization that are often found in yeast promoters, often causing disruption of nucleosome positioning (46). These are the homopolymeric stretches of deoxyadenosine present on one of the strands of double stranded DNA (46). However, these poly(dA:dT) tracts are not present in nucleosome depletion regions around TF binding sites and the cis determinants of these sites remain elusive (47). Furthermore, it is not clear if there is a sequence bias for pioneer-TF induced chromatin opening. Therefore, the current study focuses on DNA sequence patterns of nucleosomes which might play a major role in causing instability in the interactions between histones and nucleosomes at and around TF binding sites.

This research focuses on identifying the specific DNA sequence patterns at or around TF binding sites. It specifically focuses on identifying the fraction of nucleosomes with WW/SS and anti - WW/SS sequence patterns as they might be responsible for maintaining the stability of the nucleosomes. ΔNPS profiles depicting the difference between a fraction of WW/SS and antiWW/SS sequence patterned nucleosomes were created. This study tests 2 hypothesis. The first hypothesis checks if the DNA sequences, specifically anti - WW/SS patterns, are responsible for chromatin opening by pioneer transcription factors. The nucleosomes bound by
pioneer TFs seem to be intrinsically unstable, which might be due to the abundance of antiWW/SS pattern compared to WW/SS pattern. Some TF binding sites are wide open for binding. These sites do not have nucleosomes on top of that. So, the second hypothesis checks if the nucleosomes surrounding these sites tend to have more WW/SS patterns compared to anti-WW/SS patterns as these nucleosomes seem to be stable and well positioned, thereby preventing nucleosome sliding to cover TF binding sites.

The study was conducted by using CHIP-Seq data for 53 TFs in mouse embryonic stem cells (mESCs), 19 TFs in yeast. Nucleosome position data in mESCs and yeast mapped by the chemical method. These datasets were used to identify the nucleosomes that are bound to transcription factors and the nucleosome occupancy around transcription start sites (TSS) of various transcription factors was determined. The fraction of nucleosome sequences exhibiting the WW/SS sequence pattern and anti-WW/SS pattern were determined in this study. Finally, ΔNPS, that is the difference between WW/SS patterned and anti-WW/SS patterned nucleosomes, was calculated (40,41).

This study found that 32 out of 53 TFs in mESCs and 8 out of 19 TFs in yeast bound to the nucleosomes having anti-WW/SS sequence pattern which was indicated by decline in the ΔNPS values at or around TF binding sites. 21 TFs in mESCs and 11 TFs in yeast bound to wide open binding sites surrounded by well positioned nucleosomes having a WW/SS pattern that makes them stable, leaving behind the genomic sites open for TF binding which was indicated by increase in the ΔNPS values around TF binding sites. Furthermore, the study found 12 pioneer TFs in mESCs and 4 pioneer TFs in yeast that were capable of binding to nucleosomes with antiWW/SS sequence patterns that might be causing unfavorable DNA - histone interactions and inducing the pioneer TF induced chromatin opening upon binding to their target sites. This was indicated by the presence of nucleosome 0 i.e., TF binding site, and decline in ΔNPS values at or
around TF binding sites. These results indicate that the nucleosomes at or around TF binding sites tend to be unstable due to the abundance of anti-WW/SS patterns compared to WW/SS patterns in those sites resulting in chromatin opening by pioneer factors, while nucleosomes surrounding wide open TF binding sites tend to be stable due to the abundance of WW/SS patterns compared to anti-WW/SS patterns thus supporting the hypothesis.

C. Materials and Methods

Calculation of Nucleosome occupancy profile at and around the TF binding sites

Nucleosomes positioning data for the mouse embryonic stem cell (mESC) datasets mapped by chemical method was retrieved from GEO (48) and yeast from Supplementary table 2 in the Brogaard et al. study (49). The read coordinates were converted from mouse genome mm8 to mm9 using LiftOver in the UCSC genome browser. CHIP-Seq data for transcription factors like Nanog, Sox2, Oct4, Uhrf1 (GSE113915), Stat3, Esrrb (GSE97304), Zbtb2 (GSE101802), Ctcf, p300 (GSE51334), Tal1, Ctcf, Gata1 (GSE51338), c-Jun (GSE50776), Max (GSE48175), Arid3a (GSE56877), Sox2, Sox17 (GSE43275), Brg1 (GSE14344), Zic2, Otx2, Sox2, Pou5f1, Pou3f1 (GSE74636), p53 (GSE26361) were retrieved from GEO and datasets for Mafk, Hfc1, Znf384, Zc3h11a were retrieved from Mouse ENCODE.

Peak calling was done for mESCs transcription factors like Baf155, Baz1a, Brg1, Med12, Med1, Phf5a, Phrf1, Ruvbl1, Sap18, Smc3, Snf2h, Srsf1, Ssrp1, Sur2, Wstf (GSE80049), Chd1, Chd4, Chd6, Chd8, Chd9, and Ep400 (GSE64825) (48) and yeast TFs like Bye1, Bur2, Ccr4, Dhh1, Rpb7, Pcf11, Fcp1, Iws1, Not3, Cdc39, Dst1, Ctr9, Not5, Pob3, Pop2, Rpo21, Rpb3, Pta1, Ess1 (49). The FASTQ files for transcription factors were first retrieved from SRA for TFs like Chd1, Chd2, Chd4, Chd6, Chd8, Chd9, and Ep400 and European nucleotide archive (ENA) for TFs like Baf155, Baz1a, Brg1, Med12, Med1, Phf5a, Phrf1, Ruvbl1, Sap18, Smc3, Snf2h, Srsf1, Ssrp1, Sur2, Wstf. For yeast, the FASTQ files for TFs were obtained from array express
Quality control was performed for the FASTQ reads using FASTQC (51). The FASTQ files for mouse mm9 and the yeast sacCer2 reference genome were downloaded from the UCSC genome browser and the reference index was created using BWA (52).

The FASTQ reads of the TFs were aligned to the reference genome using BWA (52), replicates were merged, BAM files were generated, and their sorting was performed using samtools (53, 54). The sorted BAM files were converted to SAM files using samtools (53, 54). The tag directories for aligned SAM files and the controls were created that contain all the relevant information of the experiment in a directory using Homer which helps for CHIP-Seq analysis (55). Then the peaks were called for the aligned SAM files using findPeaks() command from the homer (55). The resulting CHIP-Seq data was used to identify the nucleosomes to which transcription factors were bound.

The CHIP-Seq peaks were first aligned with centers at position 0 i.e., the transcriptional start sites (TSSs), and the CHIP - loci range [-1000,1000] from TSSs were calculated. The nucleosome dyads within the CHIP fragments were identified and the nucleosomes were extended [-73,73] from the dyad. The average nucleosome occupancy for -1000 bp, +1000 bp relative to position 0 was calculated. Nucleosome occupancy profiles were created to get a range of nucleosomes -5, -4, ...-2, ..., 1, +2, ..., +5 based on their TF binding site i.e., position 0. Here, Nucleosome 0 is the nucleosome (147-bp) on which the center of CHIP peaks is located. Nucleosomes +/-1 are the nucleosomes that are next to the nucleosome 0. Nucleosomes +/-2 are the nucleosomes that are next to the nucleosome +/-1. Nucleosomes +/-3 are the nucleosomes that are next to the nucleosome +/-2. Nucleosomes +/-4 are the nucleosomes that are next to the nucleosome +/-3. Nucleosomes +/-5 are the nucleosomes that are next to the nucleosome +/-4.
Calculation of ΔNPS at or around TF binding sites

The phased nucleosomes, named as nucleosomes -5, -4, .... +4, and +5, were organized relative to the transcriptional start site. The nucleosomes were extended [-73,73] from the dyad. The 147bp mouse nucleosomal sequences were extracted in the FASTA format using the nucleosome dyad positions. The FASTA sequences of the mouse genome mm9 and yeast
sacCer2 were retrieved from the UCSC genome browser. These mouse genome mm9 extracted sequences were then used to do a BLAT (56) sequence search in the UCSC browser to ensure correct extraction of the sequences using the genomic coordinates. Then the forward and reverse strands of the sequences were prepared for calculating the WW (AA+TT+AT+TA) and SS (GG+CC+GC+CG) frequencies (40, 41). The number of AA, AT, TA, TT, GG, GC, CG, and CC dinucleotides was calculated in the minor and major-groove binding sites followed by the calculation of WW and SS counts on each position of the 147 bp sequences (40, 41).

Nucleosomes were further divided into four types depending on the nucleosomal sequence patterns namely type 1, type 2, type 3 and type 4 nucleosomes based on the relative abundance of WW and SS dinucleotides in the major and minor groove binding sites (41). Type 1 nucleosomes are the conventional WW/SS patterned nucleosomes defined by the abundance of WW observed in minor - groove binding sites than in major - groove binding sites and abundance of SS in major - groove binding sites than in minor - groove binding sites (41). Type 2 and type 3 nucleosomes have WW and SS abundantly present in minor-GBS or in major-GBS showing ‘mixed’ patterns (41). Type 4 nucleosome sequences have anti-WW/SS patterns where more WW dinucleotides are found in major-GBS compared to minor-GBS and more SS dinucleotides in minor-GBS compared to major-GBS (41).

ΔNPS is the difference between” Type 1 and Type 4 nucleosomes in percentage (%) is denoted as ΔNPS for a given genomic region” (41). That is, ΔNPS = Type 1 (%) – Type 4 (%) (41). ΔNPS was calculated to determine the fraction of anti - WW/SS pattern in the transcription factor bound nucleosomes and neighboring nucleosomes.
Retrieval of nucleosome positions of mESCs datasets mapped by chemical method from GEO

Peak calling using homer and retrieval of CHIP-Seq data for TFs from GEO and mouse ENCODE

Aligning CHIP-Seq peaks with centers at position 0 and get CHIP-loci range [-1000,1000] from TSSs

Identification of nucleosome dyads within the CHIP-loci range and extend the nucleosomes [-73,73] from the dyad

Calculation of average nucleosome occupancy for -1000 bp, +1000 bp relative to position 0

Creating nucleosome occupancy profiles
Collection of range of nucleosomes -1, +1, +2... +5 based on their TF binding site i.e., position 0

Identification of a range of CHIP loci and collection of nucleosome dyads in this range

Extending the nucleosomes [-73,73] from the dyad

Extraction of 147bp mouse nucleosome sequences in FASTA format from the mouse genome using dyad positions

Preparing the forward and reverse strands for the extracted sequences

Calculation of WW (AA+TT+AT+TA) and SS (GG+CC+GC+CG) frequencies in major-GBS and minor-GBS on each position of the 147 bp sequences

Determination of the fraction of 4 nucleosome sequence patterns - type1, type2, type3, type4 on each nucleosome on or near TFs

Calculation of $\Delta$NPS (Type 1 (%) – Type 4 (%))

Figure 2. Flow chart for calculation of nucleosome occupancy and $\Delta$NPS at or around the TF binding sites
The motif of a given transcription factor was obtained using the MEME suite (http://meme-suite.org/db/motifs, JASPAR CORE (2018)). This motif was used to scan the mouse genome mm9 to determine the putative binding sites of that transcription factor using the MEME FIMO module. The motifs were not available for yeast. This set of binding sites for mice was then used to determine the non-overlapping fragments between the binding sites and CHIPSeq fragments to obtain unbound sites for the given transcription factor using BEDTools (57). The number of unbound sites equal to the number of CHIP fragments of the transcription factors were randomly selected and aligned to position 0. These binding sites were extended to [1000, +1000] from position 0 and the average nucleosome occupancy for -1000 bp, +1000 bp relative to position 0 was calculated.

Mouse nucleosomal sequences were extracted within the unbound sites range and the number of WW and SS dinucleotides was calculated in the minor and major groove binding sites. The fraction of type 1 and type 4 nucleosomal sequence patterns was determined followed by calculation of ΔNPS values for the unbound motifs. This step was repeated 100 times for the randomly selected unbound sites and the mean and standard deviation of the ΔNPS was calculated for the TF and the ΔNPS profiles were created.
Aligning the unbound binding sites to position 0 and getting the range of fragments

Calculating the average nucleosome occupancy for -1000 bp, +1000 bp relative to position

Calculation of WW (AA+TT+AT+TA) and SS (GG+CC+GC+CG) frequencies in major-

Determination of the fraction of type 1 and type 4 nucleosome sequence patterns on each

Calculation of ∆NPS (Type 1 (%) – Type 4 (%))

Repeating the process multiple times (~ 100 iterations)

Calculation of mean and standard deviation for ∆NPS values for the TF unbound sites

Figure 3. Flow chart for calculation of nucleosome occupancy and ∆NPS for the TF unbound sites
**Data Overview**

**Nucleosome positions data mapped by chemical method for mESCs and yeasts**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Num</th>
<th>Accession ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse embryonic stem cells (mESCs)</td>
<td>1</td>
<td>GSM2183909</td>
<td>48</td>
</tr>
<tr>
<td>Yeasts</td>
<td>1</td>
<td>Supplementary table 2</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 1. Nucleosome positions data mapped by chemical method for mESCs and yeasts

**CHIP Seq datasets for mESCs from NCBI**

<table>
<thead>
<tr>
<th>TF</th>
<th>Num</th>
<th>GEO ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanog, Sox2, Oct4, Uhrf1</td>
<td>4</td>
<td>GSE113915</td>
<td>58</td>
</tr>
<tr>
<td>Stat3, Esrrb</td>
<td>2</td>
<td>GSE97304</td>
<td>59</td>
</tr>
<tr>
<td>Zbtb2</td>
<td>1</td>
<td>GSE101802</td>
<td>60</td>
</tr>
<tr>
<td>p300</td>
<td>1</td>
<td>GSE51334</td>
<td>61</td>
</tr>
<tr>
<td>Tal1, Gata1</td>
<td>2</td>
<td>GSE51338</td>
<td>62</td>
</tr>
<tr>
<td>c-Jun</td>
<td>1</td>
<td>GSE50776</td>
<td>63</td>
</tr>
<tr>
<td>Max</td>
<td>1</td>
<td>GSE48175</td>
<td>64</td>
</tr>
<tr>
<td>Arid3a</td>
<td>1</td>
<td>GSE56877</td>
<td>65</td>
</tr>
<tr>
<td>Sox2, Sox17</td>
<td>2</td>
<td>GSE43275</td>
<td>66</td>
</tr>
</tbody>
</table>
### Table 2. CHIP Seq datasets for mESCs from NCBI

<table>
<thead>
<tr>
<th>TF</th>
<th>Num</th>
<th>GEO ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brg1</td>
<td>1</td>
<td>GSE14344</td>
<td>67</td>
</tr>
<tr>
<td>Zic2, Otx2, Sox2, Pou5f1, Pou3f1</td>
<td>5</td>
<td>GSE74636</td>
<td>68</td>
</tr>
<tr>
<td>Mafk, Chd2, Hcfc1, Znf384, Ze3h11a</td>
<td>5</td>
<td>Mouse ENCODE</td>
<td>69</td>
</tr>
<tr>
<td>p53</td>
<td>1</td>
<td>GSE26361</td>
<td>70</td>
</tr>
<tr>
<td>Ctf, c-Myc, n-Myc</td>
<td>3</td>
<td>GSE11431</td>
<td>71</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>31</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Peak Calling using HOMER for TFs using GEO datasets in mESCs

<table>
<thead>
<tr>
<th>TF</th>
<th>Num</th>
<th>GEO ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baf155, Baz1a, Brg1, Med12, Med1, Phf5a, Phrf1, Ruvbl1, Sap18, Smc3, Snf2h, Srsf1, Ssrp1, Sur2, Wstf</td>
<td>15</td>
<td>GSE80049</td>
<td>72</td>
</tr>
<tr>
<td>Chd1, Chd2, Chd4, Chd6, Chd8, Chd9, Ep400</td>
<td>7</td>
<td>GSE64825</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>22</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Peak Calling using HOMER for TFs using GEO datasets in mESCs
Peak calling using HOMER for TFs using Array Express datasets in yeasts

<table>
<thead>
<tr>
<th>TF</th>
<th>Num</th>
<th>Array Express ID</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bye1, Bur2, Ccr4, Dhh1, Rpb7, Pcf11, Fcp1, Iws1, Not3, Cdc39, Dst1, Ctr9, Not5, Pob3, Pop2, Rpo21, Rpb3, Pta1, Ess1</td>
<td>20</td>
<td>E-MTAB-440</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 4. Peak calling using HOMER for TFs using Array Express datasets in yeasts

D. Results

The hypothesis was tested with 53 transcription factors from mouse embryonic stem cells (mESCs). The study was started with 2 pluripotent factors - Oct4 and c-Myc from mESCs - as previous studies showed that Oct4, along with Klf4 and Sox2, is able to bind nucleosomal DNA *in vitro*, whereas c-Myc does not have independent nucleosome-binding activity (74).

CHIP - seq data of Oct4 and c-Myc in which nucleosomes are mapped by the chemical method (14) based on site-directed hydroxyl radicals (75) was used from mouse embryonic stem cells. First, the 147 bp nucleosomal DNA were divided into 4 sequence patterns on the basis of WW/SS dinucleotides occurrences in major and minor - GBS. Then, the fractions of the 4 types of nucleosomes and a ΔNPS value which represents the abundance of sequence patterns either WW/SS or anti-WW/SS on the nucleosomes in a given genomic region were determined.

Nucleosome occupancy and ΔNPS profiles for TF binding sites in mESCs and yeast

Nucleosome occupancy around TSS of mouse genes was calculated along with ΔNPS values of nucleosomes -5, -4, -3, -2, ..., +2, +3, +4, +5. The study found that nucleosome occupancy peaked around the centers of the Oct4 sites but not those of the c-Myc sites (Figure 4, upper panel).
Previous in vitro studies showed that Oct 4 is able to interact with its targets in nucleosomal DNA unlike c-Myc (76). The results were found to be consistent with the previous studies.

The ΔNPS profiles for the -5, -4, -3, -2, ..., +2, +3, +4, +5 nucleosomes around the centers of the Oct4 and c-Myc binding sites revealed local minima at nucleosome 0 for Oct4 (i.e., the TFbound nucleosome) or ±1 nucleosomes for c-Myc (i.e., the nucleosomes adjacent to TF binding) (Figure 4, lower panel).

These results show that for most of the TFs like c-Myc, Oct4, Erssb, Smad1, E2f1, Baz1a, Arid3a, n-Myc, P300, Stat3, Cjun, Sox17, Sox2, Nanog, Brg1, Zbtb2, Pou3f1, Otx2, Mafk, Hcfc1, Znf384, Tcfcp2l1, Zic2, Klf4, Urhf1, Tal1, P53, Phrf1, Gata1, Max, Ep400, Med12 (Figure 4, Figure 15 - 29, lower panel (Appendix)), there is an increase in the anti-WW/SS pattern at or around TF binding sites and decrease in WW/SS pattern indicated by lower ΔNPS values at the TF

**Figure 4.** Nucleosome occupancy and ΔNPS value profiles around c-Myc (left) and Oct4 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
binding sites. As the anti-WW/SS sequence pattern is unfavorable to the DNA-histone interactions, the nucleosomes with this sequence pattern are presumably unstable. This suggests there is instability in the nucleosomes at or around TF binding sites which might facilitate nucleosome sliding or nucleosome eviction upon binding to their target sites.

This study was extended to yeast, and the hypothesis was tested in yeasts with 19 TFs. Nucleosome occupancy around TSS of yeast genes was calculated along with ΔNPS values of nucleosomes -4, -3, -2, ..., +2, +3, +4. Then, the fractions of the 4 types of nucleosomes and a ΔNPS value in a given genomic region were determined. The ΔNPS profiles for the -4, -3, -2, ..., +2, +3, +4 nucleosomes around the centers of the TF binding sites revealed local minima at nucleosome 0 for TFs like Bur2, Dhh1, Bye1, Ccr4 Rpb7, Pcf11, Fcp1, Iws1, Not3, Cdc39, Dst1, Ctr9, Not5, Pob3, Pop2, Rpo21, Rpb3 (Figure 5, Figure 39 - 41) or ±1 nucleosome for TFs like Pta1 and Ess1 (i.e., the nucleosomes adjacent to TF binding) (Figure 6, lower panel). For most of the TFs like Bye1, Bur2, Ccr4, Dhh1, Rpb7, Pcf11, Fcp1, Pta1 (Figure 6, Figure 40 - 42, lower panel), there is an increase in anti-WW/SS patterns at or around TF binding sites and decrease in WW/SS patterns indicated by lower ΔNPS values at the TF binding sites.
Figure 5. Nucleosome occupancy and ΔNPS value profiles around Bur2 (left) and Dhh1 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The decline in ΔNPS values at the sites is indicated by arrows.

Figure 6. Nucleosome occupancy and ΔNPS value profiles around Pta1 ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The decline in ΔNPS values at the sites is indicated by arrows.
Transcription factors conformations around their binding sites:

The ‘peak’ or ‘dip’ conformations of TFs around their binding sites indicate their ability to bind to nucleosomal DNA target sites. The peak in ΔNPS profile at the CHIP fragment center indicates that those TFs tend to interact with stable nucleosomes having WW/SS sequence patterns whereas a dip indicates that the TFs tend to bind to unstable nucleosomes having anti-WW/SS sequence patterns.

Peak/dip conformations of TFs around ChIP fragment center in mESCs

<table>
<thead>
<tr>
<th>TFs in mESCs</th>
<th>Number</th>
<th>Peak/Dip around TF binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-Myc, Oct4, Erssb, Smad1, E2f1, Baz1a, Arid3a, n-Myc, P300, Stat3, Cjun, Sox17, Sox2, Nanog, Brg1, Zbtb2, Pou3f1, Otx2, Mafk, Hcfc1, Znf384, Tcfcp2l1, Zic2, Klf4, Urhf1, Tal1, P53, Phrf1, Gata1, Max, Ep400, Med12</td>
<td>32</td>
<td>Dip</td>
</tr>
<tr>
<td>Ctcf, Chd2, Smc3, Srsf1, Sur2, Phf5a, Ruvbl2, Snf2h, Ssrp1, Wstf, Chd1, Chd4, Chd6, Chd8, Med1, Chd9, Sap18, Pou5fl, Zc3h11a, Zfx, Baf155</td>
<td>21</td>
<td>Peak</td>
</tr>
</tbody>
</table>

Table 5. Peak/dip conformations of TFs around ChIP fragment center in mESCs
Peak/dip conformations of TFs around ChIP fragment center in yeasts

<table>
<thead>
<tr>
<th>TFs in yeasts</th>
<th>Number</th>
<th>Peak/Dip around TF binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bye1, Bur2, Cer4, Dhh1, Rpb7, Pcf11, Fcp1, Pta1</td>
<td>8</td>
<td>Dip</td>
</tr>
<tr>
<td>Iws1, Not3, Cdc39, Dst1, Ctr9, Not5, Pob3, Pop2, Ess1, Rpo21, Rpb3</td>
<td>11</td>
<td>Peak</td>
</tr>
</tbody>
</table>

Table 6: Peak/dip conformations of TFs around ChIP fragment center in yeasts

Peak/dip conformations of TFs having nucleosome 0/without nucleosome 0 around their binding sites in mESCs

<table>
<thead>
<tr>
<th>TFs in mESCs with nucleosome 0</th>
<th>Peak/Dip around TF binding sites</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct4, Smad1, Baz1a, Arid3a, P300, c-Jun, Sox17, Sox2, Nanog, Brg1, Otx2, Ep400</td>
<td>Dip</td>
<td>12</td>
</tr>
<tr>
<td>Ssrp1, Chd4, Chd6, Med1, Chd8, Chd9, Baf155, Chd2</td>
<td>Peak</td>
<td>8</td>
</tr>
</tbody>
</table>
Table 7: Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in mESCs

<table>
<thead>
<tr>
<th>TFs in mESCs without nucleosome 0</th>
<th>TFs in mESCs with nucleosome 0</th>
<th>Peak/Dip around TF binding sites</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>TF</td>
<td>Oct4, Sox2</td>
<td>Dip</td>
</tr>
<tr>
<td>c-Myc, Erssb, E2f1, n-Myc, Stat3,</td>
<td></td>
<td>Klf4, p53</td>
<td>Dip</td>
</tr>
<tr>
<td>Zbtb2, Pou3f1, Mafk, Hcfc1,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tcfcp2l1, Znf384, Zic2, Klf4,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urhf1, Tal1, P53, Phrf1, Gata1,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max, Med12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctcf, Smc3, Srsf1, Sur2, Phf5a,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snf2h, Ruvbl1, Wstf, Chd1, Sap18,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pou5f1, Zc3h11a, Zfx</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in mESCs for validated pioneer factors

TFs in mESCs with nucleosome 0

<table>
<thead>
<tr>
<th>TF</th>
<th>Peak/Dip around TF binding sites</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct4, Sox2</td>
<td>Dip</td>
<td>2</td>
</tr>
</tbody>
</table>

TFs in mESCs without nucleosome 0

<table>
<thead>
<tr>
<th>TF</th>
<th>Peak/Dip around TF binding sites</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klf4, p53</td>
<td>Dip</td>
<td>2</td>
</tr>
</tbody>
</table>

Pou5f1                              | Peak                             | 1      |
Table 8: Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in mESCs for validated pioneer factors (57)

**Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in yeasts**

<table>
<thead>
<tr>
<th>TFs in yeasts with nucleosome 0</th>
<th>Peak/Dip around</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bur2, Dhh1, Bye1, Ccr4, Rpb7, Pcf11, Fcp1</td>
<td>Dip</td>
<td>7</td>
</tr>
<tr>
<td>Cdc39, Not3, Rpo21, Iws1, Dst1, Ctr9, Not5, Pob3, Pop2, Rpb3</td>
<td>Peak</td>
<td>10</td>
</tr>
<tr>
<td>Ess1</td>
<td>Dip</td>
<td>1</td>
</tr>
<tr>
<td>Pta1</td>
<td>Peak</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 9: Peak/dip conformations of TFs having nucleosome 0/ without nucleosome 0 around their binding sites in yeasts

**Nucleosome occupancy and ΔNPS profiles for TF binding sites in mESCs and yeast with increase of ΔNPS around TF binding sites**

It was observed that majority of the nucleosomes at or around TFs in mESCS and yeast to be having Anti – WW/SS sequence patterns. However, in mESCs some nucleosomes are found to have stable structure and are found to be well organized around TF binding sites. These are indicated by the peaks at the center in their ΔNPS profiles of TFs like CTCF (both chemical and MNase), Chd2, Smc3, Srsf1, Sur2, Phf5a, Ruvbl2, Snf2h, Ssrp1, Wstf, Chd1, Chd4, Chd6, Chd8,
Med1, Chd9, Sap18, Pou5f1, Zc3h11a, Zfx, Baf155 (Figure 7, Figure 30 - 39, lower panel (Appendix)). These results suggest that these nucleosomes tend to have more WW/SS sequence patterns than anti-WW/SS sequence patterns at these TF binding sites. This indicates that these nucleosomes are intrinsically stable, which leaves the binding sites wide open that will be favorable for TF binding.

In yeast, peaks observed at the center of ΔNPS profiles for TFs like Iws1, Not3, Cdc39, Dst1, Ctr9, Not5, Pob3, Pop2, Ess1, Rpo21, Rpb3 (Figure 8 - left, Figure – 9, Figure 39- 42, lower panel) show similar profiles as mESC TFs. These results suggest that few nucleosomes
tend to have more anti-WW/SS sequence patterns than WW/SS sequence patterns at TF binding sites. This indicates that the nucleosomes are intrinsically unstable or stable depending on the presence of the WW/SS and anti-WW/SS sequence patterns.

Figure 8. Nucleosome occupancy and ΔNPS value profiles around Ess1 (left) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.
Most TFs (33 out of 53) do not have a nucleosome 0 which is consistent with nucleosome depleted regions (NDR) around TF binding sites. Out of 33 TFs, 20 TFs were found to have the dip profile, indicating that anti-WW/SS is more abundant than WW/SS pattern. So, we speculate that anti-WW/SS patterns may be associated with the formation of NDR.

For TFs with nucleosome 0 (20 out of 53), most of them (12 out 20) have the ‘dip’ profile and interestingly, these TFs include several well-known pioneer factors such as Oct4 and Sox2. Therefore, we speculate that the anti-WW/SS pattern may contribute to the chromatin opening by pioneer factors. Based on the results obtained from the study, two schematic models were created.

The first model represents the TFs binding to their target sites that are wide open for binding without nucleosomes in the binding site. The neighboring nucleosomes -3, -2, -1 and +1, +2, +3 tend to have more WW/SS sequence patterns making them stable compared to nucleosomes with

---

**Figure 9. Nucleosome occupancy and ΔNPS value profiles around Cdc39 (left) and Not3 (right) ChIP sites.** Nucleosomes -3 to +3 (left) and nucleosomes -2 to +2 (right) around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.
anti-WW/SS sequence patterns. These nucleosomes are well positioned, thereby preventing the nucleosome sliding to cover the TF binding sites (Figure 10).

![Nucleosomes with WW/SS sequence patterns](image)

**Figure 10. TF binding activity to their target sites wide open surrounded by stable nucleosomes having WW/SS sequence patterns**

The second model represents independent binding activity of pioneer factors to their target sites on the nucleosomes having anti-WW/SS sequence patterns. These pioneer factors are capable of binding to such nucleosomes as they will be intrinsically unstable. The instability of the nucleosomes may be due to the abundance of anti-WW/SS patterns in those sites. This will result in chromatin opening by pioneer factors resulting in nucleosome eviction (Figure 11).
Figure 11. Independent binding activity by pioneer factors to their target sites on anti-WW/SS nucleosomes
Nucleosome occupancy and ΔNPS profiles for TF unbound sites in mESCs

The unbound sites of selected TF like Arid3a in mice for which data was available were obtained using MEME and bedtools. The unbound sites were randomly selected, and nucleosome occupancy was calculated for nucleosomes -4, -3, -2, ..., +2, +3, +4. Then, the fractions of the 4 types of nucleosomes and a ΔNPS value in a given genomic region were determined. This was performed for the number of iterations equal to the number of CHIP-Seq peaks for the TF followed by calculation of mean, standard deviation and p-values for the ΔNPS values. One-sample t-test was performed to determine the statistical significance of the results. The null hypothesis assumes no difference in the ΔNPS values between the bound and unbound sites. The alternative hypothesis assumes that the ΔNPS values of bound sites is different from unbound sites. The empirical p-values were found to be < 0.05 rejecting the null hypothesis (Table 10) thus, supporting the hypothesis of the study. The ΔNPS profile depicting both mean and standard deviation of ΔNPS values was created for the TF unbound sites. It was observed that there is no increase or decrease in the fraction of WW/SS and Anti-WW/SS sequence patterns at or around the TF unbound sites. These results suggest that these types of nucleosomal sequence patterns are only present in abundance at or around the TF binding sites.
Figure 12. ∆NPS value profiles around Arid3a unbound sites indicated by blue and Arid3a binding sites indicated by red line. The decline in ∆NPS values at the binding sites is indicated by arrows.

<table>
<thead>
<tr>
<th>Nucleosome</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>8.5512e-08</td>
</tr>
<tr>
<td>-3</td>
<td>1.0074e-07</td>
</tr>
<tr>
<td>-2</td>
<td>7.5904e-07</td>
</tr>
<tr>
<td>-1</td>
<td>1.6284e-07</td>
</tr>
<tr>
<td>0</td>
<td>2.0354e-08</td>
</tr>
<tr>
<td>1</td>
<td>1.0828e-06</td>
</tr>
<tr>
<td>2</td>
<td>3.6996e-07</td>
</tr>
</tbody>
</table>
Table 10: p-values obtained from one-sample t-test showing statistical significance of differences in the ΔNPS values around TF bound and unbound sites

E. Discussion

The objective of this study was to determine the nucleosomal DNA sequence patterns at and around TF binding sites that might be responsible for forming nucleosome depleted regions and pioneer TF induced chromatin openings.

The study was conducted on 53 TFs from mESCs and 20 TFs from yeast. The results showed that 32 out of 53 TFs in mESCs and 8 out of 19 TFs in yeast had lower ΔNPS values around their binding sites. ΔNPS profiles of mESC TFs like Sox2, Oct4, c-Myc Nanog, Pou3f1, p53 and, n-Myc and yeast TFs like Cdc39, Not3, Bur2, Rpb7, Pcf11 and, Fcp1 show a definite ‘dip’ at the CHIP fragment center. This indicated an increase in the anti-WW/SS pattern at or around TF binding sites that could be responsible for causing instability in the region by affecting the nucleosome - histone interactions and resulting in nucleosome eviction.

TFs like Oct4, Sox2, Nanog, Smad1, Baz1a, Arid3a, P300, Sox17, c-Jun, Brg1, Otx2 and, Ep400 in mESCs, and TFs like Bur2, Pcf11, Rpb7, Ccr4 and, Fcp1 in yeast show nucleosome occupancy at their binding sites i.e., nucleosome 0, and show a decline in ΔNPS profiles at their binding sites; suggesting that these TFs do not compete for their binding sites with nucleosomes. These TFs act as pioneer factors resulting in induced chromatin opening which might be caused by instability in the region because of the presence of anti-WW/SS
nucleosomes. This suggests that these TFs are capable of binding independently to the nucleosomal DNA and help in recruiting TFs that do not show independent binding activity to their target sites in the nucleosomal DNA.

However, the remaining 21 TFs in mESCs and 11 TFs in yeast had an increase in ΔNPS at the TF binding sites. ΔNPS profiles showed definite ‘peaks’ at the center of CHIP fragments between nucleosomes -3 to +3 for certain mESCs TFs like Ctcf, Srsf1, Phf5a, Wstf, Chd4, and Chd9 and between nucleosomes -2 and +2 for yeast TFs like Ccr4, Dhh1, Iws1, Dst1, Ctr9, Not5 and, Rpb3. Particularly, Ctcf binding in mESCs showed ‘peak’ in ΔNPS profile at CHIP fragments center in nucleosome mapping in both MNase and chemical method. This indicated the presence of WW/SS patterns abundantly present on the nucleosomes surrounding their target sites suggesting that these nucleosomes are intrinsically stable leaving the binding sites wide open that will be favorable for TF binding. This shows the well-positioning of the nucleosomes -4, -3, -2, -1, +1, +2, +3, +4 that prevents nucleosome sliding to cover Ctcf binding sites.

Also, ΔNPS profiles using both chemical and MNase data was created only for Ctcf. This is because Ctcf has regular nucleosome arrays around the binding sites. For this particular TF, good boundaries for nucleosome -3, -2, -1, +1, +2, +3 etc. were obtained, while MNaseseq data for other TFs do not have well-defined regions for nucleosomes. So, it becomes difficult to compare delta NPS profiles.

This study also shows that the nucleosomes in mESCs tend to have abundant antiWW/SS sequence patterns at and around TF binding sites compared to WW/SS sequence patterned nucleosomes. Conversely, in yeast WW/SS sequence patterned nucleosomes are abundantly observed. Therefore, we see differences in the ΔNPS profiles between mESCs and yeast.
This might be due to the increased transcriptional levels in mammals as compared to non-mammalian genes as observed in previous studies (64, 77).

Additionally, ΔNPS values at the TF unbound sites for Arid3a were calculated. The hypothesis was supported by a ΔNPS profile created which showed that there is no rise or decline in the ΔNPS values at the TF unbound sites in mESCs.

Results obtained from this study suggest that certain TFs can bind to their target sites without competing with nucleosomes. These include TFs whose centers show nucleosome occupancy i.e., nucleosome 0. Conversely, certain TFs bind to nucleosomal DNA by competing with them, causing a nucleosome depletion region formation adjacent to TF binding sites (±1 nucleosomes). The nucleosomes are observed to have anti-WW/SS sequence patterns at these sites, and they seem to be causing instability between histone – nucleosome interactions.

Altogether, the results suggest that WW/SS and Anti-WW/SS types of nucleosomal sequence patterns are present in abundance only at or around the TF binding sites. Particularly, nucleosomes with anti-WW/SS sequence patterns might be responsible for the chromatin induced opening by pioneer factors by causing instability in the region. Further, this research might help in developing methods to control the expression of certain TFs as this study focuses on determining the molecular mechanism causing the chromatin opening by pioneer TFs which might be applicable for cancer studies.
F. Future work

In the current study, nucleosome occupancy and ΔNPS profiles in mESCs and yeast mapped by chemical method were determined. The analysis was performed for MNase-Seq datasets in mESCs as well. However, it was observed that for most TFs, except Ctcf, MNase-seq data give poor boundaries (Figure 9 – 10), represented by nucleosome occupancy and ΔNPS profiles for Oct4 and Nanog mapped by MNase-seq data. Therefore, nucleosome occupancy and ΔNPS profiles for MNase-seq data should be recalculated with the nucleosome ranges (i.e., nucleosome +/-3, +/-2, +/-1) derived from chemical methods. Generally, it is assumed that both the methods should be giving the same ranges for nucleosome +/-3, +/-2, and +/-1 if they are equally accurate. However, if the boundaries remain poor as the nucleosomes are not regularly positioned in MNase-seq data, that might be due to underrepresentation of nucleosomes with sequences that are A/T rich (14), boundaries defined by chemical method could be used instead (77). The current model proposed in this study can be further examined by performing statistical colocalization analysis like Fisher’s exact test to validate the conclusions. Another application for the study is to determine the mechanism of nucleosome positioning using the existing deformation energy model (78). The model selected should consider the bending of DNA (78) to predict nucleosome positions to identify if the nucleosome WW/SS and Anti – WW/SS sequence patterns plays a role in nucleosome organization.
Figure 13. Nucleosome occupancy and $\Delta$NPS value profiles around Oct4 ChIP sites – Chemical method (left) and MNase-Seq method (right). Nucleosomes -3 to +3 (left) and nucleosomes -2 to +2 (right) around the binding sites are demarcated by dashed lines. The increase in $\Delta$NPS values at the sites is indicated by arrows.

Figure 14. Nucleosome occupancy and $\Delta$NPS value profiles around Oct4 ChIP sites – Chemical method (left) and MNase-Seq method (right). Nucleosomes -3 to +3 (left) and nucleosomes -2 to +2 (right) around the binding sites are demarcated by dashed lines. The increase in $\Delta$NPS values at the sites is indicated by arrows.
G. Literature Cited


Genome Res. 24, 1945-1962.


genome-wide binding sites analysis links the Myc regulatory network with the polycomb

and Kim, J. (2014) Arid3a is essential to execution of the first cell fate decision via direct

66. Aksoy, I., Jauch, R., Chen, J., Dyla, M., Divakar, U., Bogu, G.K., Teo, R., Leng Ng, C.K.,
Oct4 switches partnering form Sox2 to Sox17 to reinterpret the enhancer code and specify
endoderm. EMBO J. 32, 938-953.

stem cell chromatin remodeling complex, esBAF, is an essential component of the core

68. Matsuda, K., Mikami, T., Oki, S., Iida, H., Andrabi, M., Boss, J.M., Yamaguchi, K.,
five major transcription factors highlights a central role for ZIC2 in the mouse epiblast

69. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S.,
Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C.,


H. Supplemental Materials

Nucleosome occupancy and ΔNPS profiles for TF binding sites in mESCs

Figure 15. Nucleosome occupancy and ΔNPS value profiles around Erssb (left) and Smad1 (right) ChIP sites. Nucleosomes -5 to +5 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.

Figure 16. Nucleosome occupancy and ΔNPS value profiles around E2f1 (left) and Baz1a (right) ChIP sites. Nucleosomes -5 to +5 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
Figure 17. Nucleosome occupancy and ΔNPS value profiles around Arid3a (left) and n-Myc (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.

Figure 18. Nucleosome occupancy and ΔNPS value profiles around P300 (left) and Stat3 (right) ChIP sites. Nucleosomes -3 to +3 (left) and nucleosomes -4 to +4 (right) around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
Figure 19. Nucleosome occupancy and ∆NPS value profiles around C-jun (left) and Sox17 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The decline of ∆NPS values at the sites is indicated by arrows.

Figure 20. Nucleosome occupancy and ∆NPS value profiles around Sox2 (left) and Nanog (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ∆NPS values at the sites is indicated by arrows.
Figure 21. Nucleosome occupancy and ∆NPS value profiles around Brg1 (left) and Zbtb2 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The decline of ∆NPS values at the sites is indicated by arrows.

Figure 22. Nucleosome occupancy and ∆NPS value profiles around Pou3f1 (left) and Otx2 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The decline of ∆NPS values at the sites is indicated by arrows.
Figure 23. Nucleosome occupancy and ΔNPS value profiles around Mafk (left) and Hcfc1 (right) ChIP sites. Nucleosomes -5 to +5 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.

Figure 24. Nucleosome occupancy and ΔNPS value profiles around Znf384 (left) and Tcfcp2l1 (right) ChIP sites. Nucleosomes -5 to +5 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
Figure 25. Nucleosome occupancy and ΔNPS value profiles around Zic2 (left) and Klf4 (right) ChIP sites. Nucleosomes -5 to +5 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.

Figure 26. Nucleosome occupancy and ΔNPS value profiles around Urhf1 (left) and Tal1 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
Figure 27. Nucleosome occupancy and ΔNPS value profiles around P53 (left) and Phrfl (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.

Figure 28. Nucleosome occupancy and ΔNPS value profiles around Gata1 (left) and Max (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ΔNPS values at the sites is indicated by arrows.
Figure 29. Nucleosome occupancy and ∆NPS value profiles around Ep400 (left) and Med12 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The decline of ∆NPS values at the sites is indicated by arrows.

Figure 30. Nucleosome occupancy and ∆NPS value profiles around Smc3 (left) and Srsf1 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.
Figure 31. Nucleosome occupancy and ∆NPS value profiles around Sur2 (left) and Phf5a (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.

Figure 32. Nucleosome occupancy and ∆NPS value profiles around Ruvbl1 (left) and Snf2h (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.
Figure 33. Nucleosome occupancy and ΔNPS value profiles around Ssrp1 (left) and Wstf (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.

Figure 34. Nucleosome occupancy and ΔNPS value profiles around Chd1 (left) and Chd4 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.
Figure 35. Nucleosome occupancy and ΔNPS value profiles around Chd6 (left) and Chd8 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.

Figure 36. Nucleosome occupancy and ΔNPS value profiles around Med1 (left) and Chd9 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.
Figure 37. Nucleosome occupancy and ∆NPS value profiles around Sap18 (left) and Pou5f1 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.

Figure 38. Nucleosome occupancy and ∆NPS value profiles around Zc3h11a (left) and Zfx (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.
Figure 39. Nucleosome occupancy and ∆NPS value profiles around Baf155 (left) and Chd2 (right) ChIP sites. Nucleosomes -4 to +4 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.
Nucleosome occupancy and ∆NPS profiles for TF binding sites in yeasts

**Figure 40.** Nucleosome occupancy and ∆NPS value profiles around Bye1 (left) and Ccr4 (right) ChIP sites. Nucleosomes -2 to +2 around the binding sites are demarcated by dashed lines. The decline in ∆NPS values at the sites is indicated by arrows.

**Figure 41.** Nucleosome occupancy and ∆NPS value profiles around Rpb7 (left) and Pcf11 (right) ChIP sites. Nucleosomes -2 to +2 around the binding sites are demarcated by dashed lines. The decline in ∆NPS values at the sites is indicated by arrows.
Figure 42. Nucleosome occupancy and $\Delta$NPS value profiles around Fcp1 (left) and Rpo21 (right) ChIP sites. Nucleosomes $-2$ to $+2$ and nucleosomes $-3$ to $+3$ around the binding sites are demarcated by dashed lines. The decline in $\Delta$NPS values at the sites is indicated by arrows.

Figure 43. Nucleosome occupancy and $\Delta$NPS value profiles around Iws1 (left) and Dst1 (right) ChIP sites. Nucleosomes $-2$ to $+2$ around the binding sites are demarcated by dashed lines. The increase in $\Delta$NPS values at the sites is indicated by arrows.
Figure 44. Nucleosome occupancy and ΔNPS value profiles around Ctr9 (left) and Not5 (right) ChIP sites. Nucleosomes -2 to +2 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.

Figure 45. Nucleosome occupancy and ΔNPS value profiles around Pob3 (left) and Pop2 (right) ChIP sites. Nucleosomes -3 to +3 around the binding sites are demarcated by dashed lines. The increase in ΔNPS values at the sites is indicated by arrows.
Figure 46. Nucleosome occupancy and ∆NPS value profiles around Rpb3 ChIP sites. Nucleosomes -2 to +2 around the binding sites are demarcated by dashed lines. The increase in ∆NPS values at the sites is indicated by arrows.