UNIVERSITY OF CALIFORNIA

Los Angeles

# Statistical Analyses of Clustering Patterns of
# Transcription Factor-DNA Binding in ChIP-seq Data

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Statistics

by

Jun Liu

2014

# ABSTRACT OF THE THESIS

Statistical Analyses of Clustering Patterns of

Transcription Factor-DNA Binding in ChIP-seq Data

by

Jun Liu

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Qing Zhou, Chair

Binding of transcription factors on specific sites of DNA is central to the regulation of gene expression. ChIP-seq technology is a novel tool that combines the method of chromatin immunoprecipitation (ChIP) with the next generation DNA sequencing (seq) to identify the transcription factor binding loci on DNA. ChIP-seq has revolutionized the process of biological data acquisition for elucidating fundamental gene regulation mechanisms. However, the acquired large dataset on transcription factor-DNA binding calls for analyses using statistical tools, which will provide predictions that guide the wet-lab biological research. This research is part of statistical modeling of patterns of transcription factor-DNA binding which serves to analyze the various patterns of transcription factor co-clustering on DNA in a ChIP-seq dataset obtained in the mouse embryonic stem cells for 15 transcription factors/coregulators. First, we used the Chi-square goodness of fit test to determine whether the location of binding sites for each transcription factor constitute a Poisson process. The results indicated that it is unlikely to be a homogenous Poisson process. Second, we studied the correlation among the bindings by various transcription factors. Third, the patterns of various clustered sites containing three transcription factors were analyzed. It is found that there are a total of 3353 such sites. The transcription factors Smad1, Tcfcp2l1, Stat3, Klf4 and Esrrb and the coregulator p300 are preferentially co-localized with Nanog, Oct4, Sox2, while E2f1 and Zfx are preferentially colocalized with n-Myc and c-Myc.

The thesis of Jun Liu is approved.

Robert L. Gould

Yingnian Wu

Qing Zhou, Committee Chair

University of California, Los Angeles

2014

## ACKNOWLEDGEMENT

I would like to thank all my committee members, Dr. Qing Zhou, Yingnian Wu and Robert Gould. Their teaching efforts in the statistics courses I took have helped transform me into a statistician. I would also like to thank my classmates in the master program of UCLA Statistics and the PhD students who are my friends, TAs and instructors in the courses I have taken. Days and nights we have spent a lot of good time together enjoying attending class, solving homework problems, or simply hanging out in and outside of school. I would especially like to thank Albert Wong, who showed me how to do rigorous mathematical proof in the canonical way.

I have received the traditional trainings in biology, in which my main activities were performing wet lab biological experiments on the bench. However, the data collection process is in general expensive and time-consuming while the data is mostly descriptive and not rigorous. One observation I made for biological data is that the data are highly variable, making one wonder whether the biological entity to be revealed by the data is best described using random variables, necessitating the need for every biologist to be a statistician. With the biotechnological advances bringing down the cost of collecting data, I am glad I am in graduate school studying statistics and thankful to Prof. Qing Zhou for the ChIP-seq data analysis project, which gives me a chance to deduce statistical significance and biological meanings in the data.

## Table of Contents

## List of Figures

## List of Tables

.

# CHAPTER 1. BACKGROUND AND GOAL

## 1.1. Transcription factor bindings and their regulation of gene expression

This section provides a brief introduction to transcription factor-DNA binding, as the thesis work is focused on the statistical analysis of transcription factor-DNA binding. Transcription factor binding to specific regions of the genomes is the most important control mechanism regulating gene expression. Such control is the molecular basis for the response of any living organism to changes in the environments. The specific regions of the genomes, called *cis*-regulatory elements, include promoters, enhancers, and silencers and binding of transcription factor to these regions inhibits or stimulates the activity of the basal transcription machinery that performs the task of genes transcription, i.e., flow of genetic information from DNA to RNA (Hobert 2008) (Figure 1).

It has been noted that the number of genes in an organism far outnumbers those of transcription factors. For instance, the human genome has about 20,000–25,000 genes with diverse and unique spatial and temporal patterns of expression, while the number of transcription factors is only about 1850 (Venter et al. 2001). Therefore, one question that arises is how a eukaryotic cell exhibits diverse gene expression profiles in response to almost limitless external signals using such a limited set of transcription factors. One solution is the combinatorial control of transcription factors (Reményi et al. 2004), which has at least two aspects. One is that the transcription factors of the same and different type can combine to form dimers or multimers to regulate transcription. For example, N types of transcription factor can combine to form a total of $N^2$ types of dimers including heterodimers or homodimers, which greatly expand the regulatory capability. One well-understood example is the formation of Oct4 and Sox2 heterodimers in embryonic stem cells that operate in later stages of embryogenesis to facilitate eye development. Oct4 and Sox2 are each capable of forming a network of transcription factor complex together with their interacting partners and regulating diverse genes (Reményi et al. 2004) (Figure 2). The second is the coregulators, which are proteins that cannot directly bind DNA but form complex with transcription factor and affects transcription by either stabilizing or destabilizing interactions between transcription factor and the basal transcription machinery. Recruitment of coregulators renders the transcription factor with considerable increased functional flexibility (McKenna and O'Malley 2002). By the different permutations of transcription factors and coregulators, even a small number of transcription factors binding to a limited number of cis-regulatory elements form a system capable of inducing the innumerable gene expression profiles.

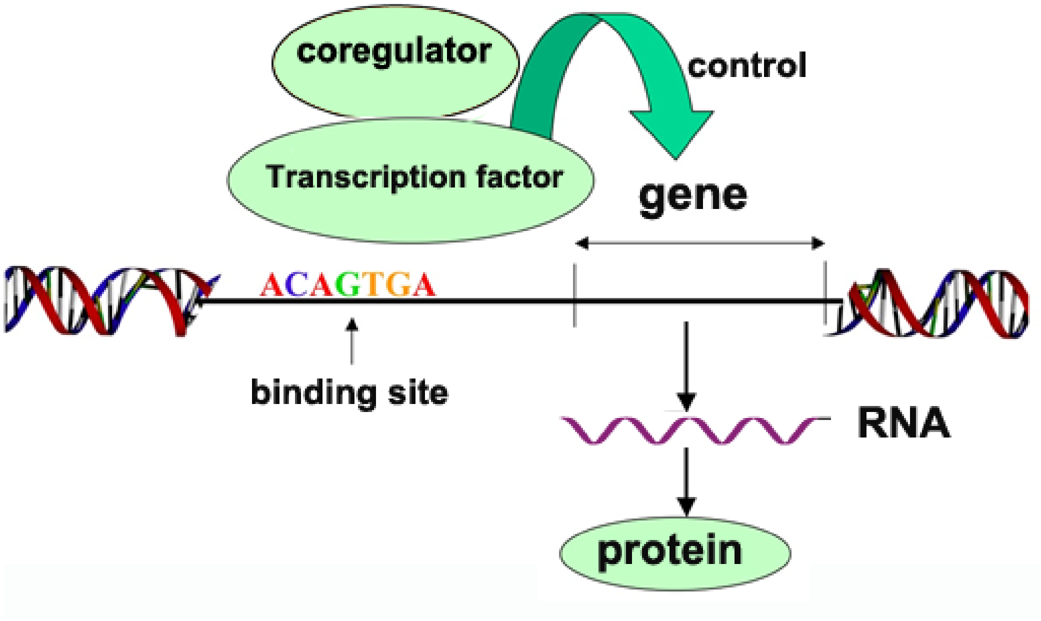


**Figure 1.** Control of gene expression by transcription factors.

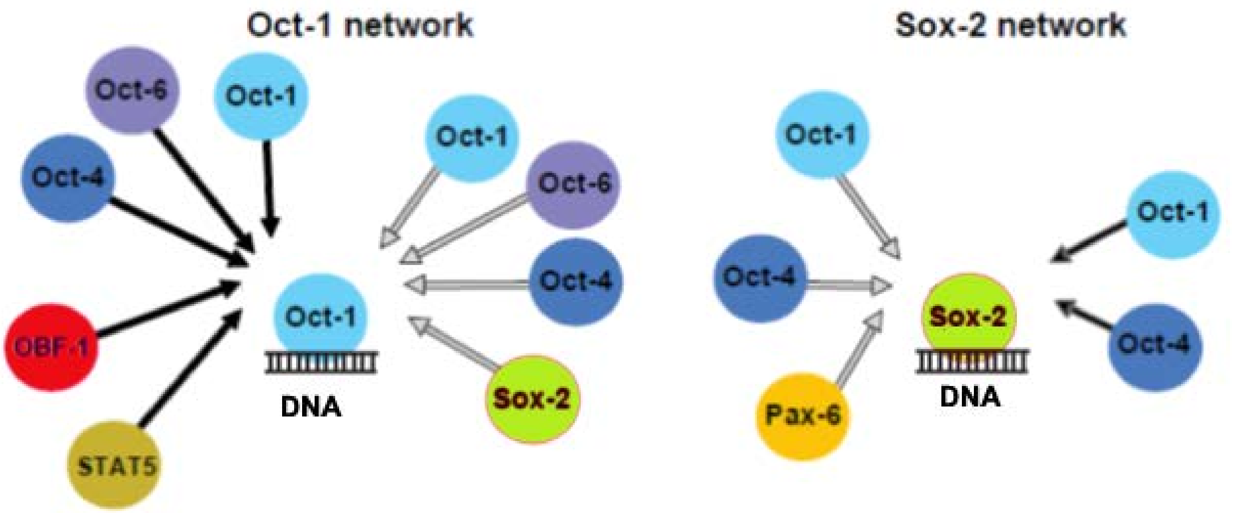


**Figure 2.** Interaction network of transcription factor Oct-1 and Sox-2. Upon binding to a specific DNA region, transcription factors Oct-1 or Sox-2 can interact with diverse partners of transcription factors.

## 1.2. ChIP-seq

Despite the fact that transcription factor and DNA binding has been studied using wet-lab technique for decades, recent advances in high-throughput next-generation DNA sequencing technology has caused the emergence of chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) technique that have revolutionized the field of transcription factor/DNA binding, a foundation of biology (Barski et al. 2007; Johnson et al. 2007). ChIP-seq research normally creates large datasets that exceed the traditional analytical capability in web-lab experimental investigation and necessitates the usage of statistical tools to make sense out of the large dataset, which is the focus of this thesis work. Before I delve into the analysis of

ChIP-seq data, it is necessary to briefly introduce the acquisition process for better understanding of the meaning of the ChIP-seq data. ChIP-seq is derived from ChIP-chip (Johnson et al. 2007), which uses hybridization-based microarray, to reveal the identities of DNA sequence among of fixed number of DNA fragments on the microarray chip that interacts with proteins. With the rapid advance of next generation sequencing (NGS) technology, ChIP-seq quickly replaced ChIP-chip and directly decipher the specific and exact sequence of DNA involved in interaction of proteins, with higher resolution and lower cost. The data acquisition process is as follows in the work flow (Figure 3).

1) The mouse embryonic stem cells are lysed and the genomic DNA with the bound transcription factors is sheared into small fragments using sonication or hydrolysis enzymes. Binding of the transcription factors to the specific regions of genomic DNA protects the DNA from being degraded by sonication and the hydrolysis enzymes and the sequence of the specific DNA fragment can be determined in subsequent steps.

2) Then antibodies for transcription factors are added to the cells lysate. The antibody binds and precipitates its cognate transcription factors, which brings down the DNA fragment bound to the transcription factors from the solution. This process is called chromosome immunoprecipitation. It is necessary that high quality antibodies for specifically precipitating the transcription factors be used for a success ChIP-seq outcome (Davies et al. 2011).

3) DNA fragments bound to the transcription factors and the antibody are recovered from the lysate and sequenced via massive parallel short-tag-based sequencing to reveal their exact sequences. The most important issue to be considered is the sequencing depth , which is defined as “The number of reads for each base” (Kunin et al. 2008). Too low

depth will lose accuracy while too high depth will incur unnecessary cost. To control for the appropriate sequencing depth, appropriate amount of sample is needed. This ends the wet-lab steps.

4) To process the raw sequence data to identify the binding peaks, two main statistical algorithms have been developed, tag shift and peak extension, used in at least 11 publicly available programs (Wilbanks and Facciotti 2010). In the report (Chen et al. 2008) which generated the dataset for being used in the thesis, the DNA sequences was analyzed by using extended 200 bp to the 3' direction. Even though this step utilizes intensively statistical analysis, the algorithms have been rather mature and are not the focus of the thesis research.

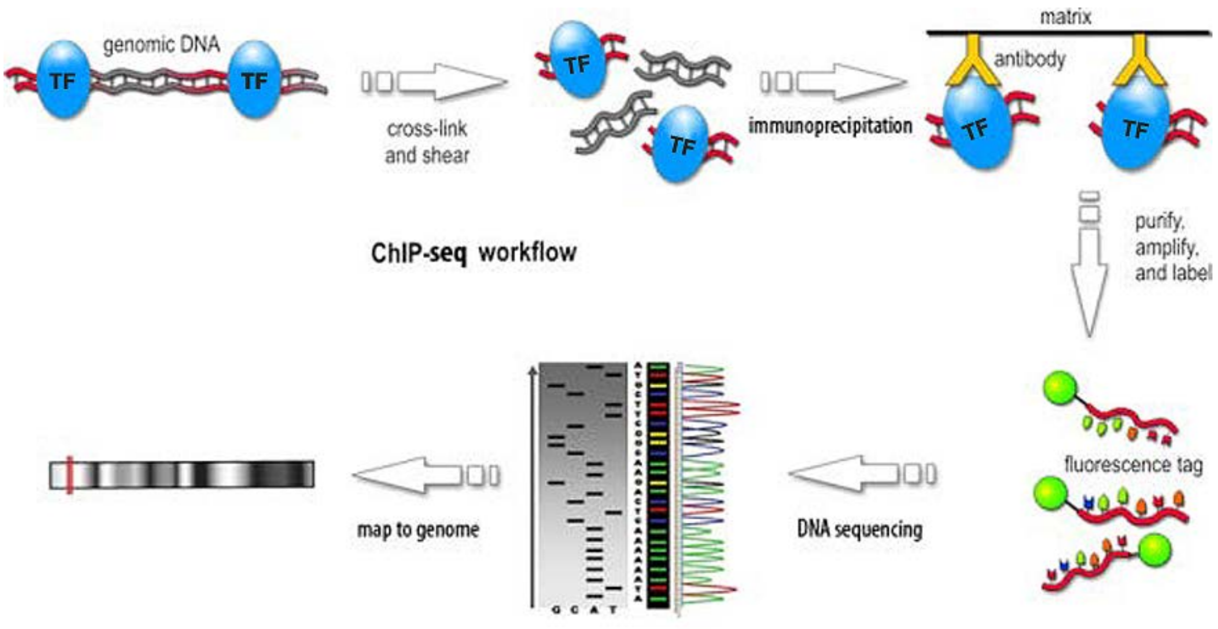


**Figure 3.** A schematic of the workflow of ChIP-seq process for deducing the binding site in the genome for transcription factors.

### 1.3. Transcription factor-DNA binding in mouse embryonic stem cells

The ChIP-seq dataset for the thesis research is from a recent report by Chen et al. which mapped the binding sites of the 15 transcription factors and coregulators in the genome of the mouse embryonic stem cells (Chen et al. 2008). This is a paper of high impact with more than 1078 citation as December 9, 2013. The importance of the report is primarily due to two reasons. First, the mouse embryonic stem cells, isolated from mouse blastocysts, are a stem cell type with potential to be differentiated to any cell types having both biological and practical importance. Biologically, mouse embryonic stem cells have been used as a model system to study the fundamental biological issue of maintenance of the self-renewing state of stem cells regulated by transcription factors (Zhou et al. 2007; Young 2011). Practically, the results contributed to stem cell research, which has been one of the hottest research areas in recent years with the potentials of stem cells to differentiate into any cell types and organs and hopes for curing diverse diseases. The finding that four transcription factors alone, Oct4, Sox2, c-Myc and Klf4, transferred into mouse fibroblasts can turn the fibroblast into a stem cell-like cells capable of differentiating into any cell types (Takahashi and Yamanaka 2006) resulted in the Novel prize awarding to Yamanaka. Recent research has found that 2 transcription factor, NOS and Lin8, transferred are enough to turn a fibroblast into stem cell-like cells (Yu et al. 2007). Secondly, it was among the seminal papers that study in a genome wide scale the various transcription factors and coregulators using ChIP-seq in mouse embryonic stem cells. The major finding of the paper was that in mouse embryonic stem cells the transcription factors are attached to the genome in two major clusters. The first is that Nanog, Oct4, Sox2, Smad1, and STAT3 form a complex. The second is that c-Myc, n-Myc, Zfx, and E2f1 form another complex. In addition, it was found in follow-up wet-lab experiments that the first clusters of transcription factors bound regions of

DNA that are an enhancers to form enhanceosomes, which is a nuclear structure composed of the complex of transcription factors and enhancers to enhance gene expression (Chen et al. 2008). Despite the 76 pages total for the text and figures plus more tables of processed ChIP-seq data, the report still barely scratched the surface of the profound and fundamental issues of transcription factor-DNA binding in the highly anticipated mouse embryonic stem cells. Despite the fact the solving the transcription factor-DNA binding enigma requires the combined effort of wet-lab biologist and statisticians, due to the size of the large dataset, even many statistical aspects remain to be tapped into, which is the focus of the current thesis research. For example, the second complex composed of c-Myc, n-Myc, Zfx and E2f1 have not been fully analyzed beyond the facts that they form clusters in a correlation analysis. And there is no spatial statistical modeling for the genome wide binding of transcription factor with DNA. In addition, transcription factor binding pattern can be associated with the expression data of nearby genes from the wet-lab experiments, to reveal gene ontology. These points will be address below.

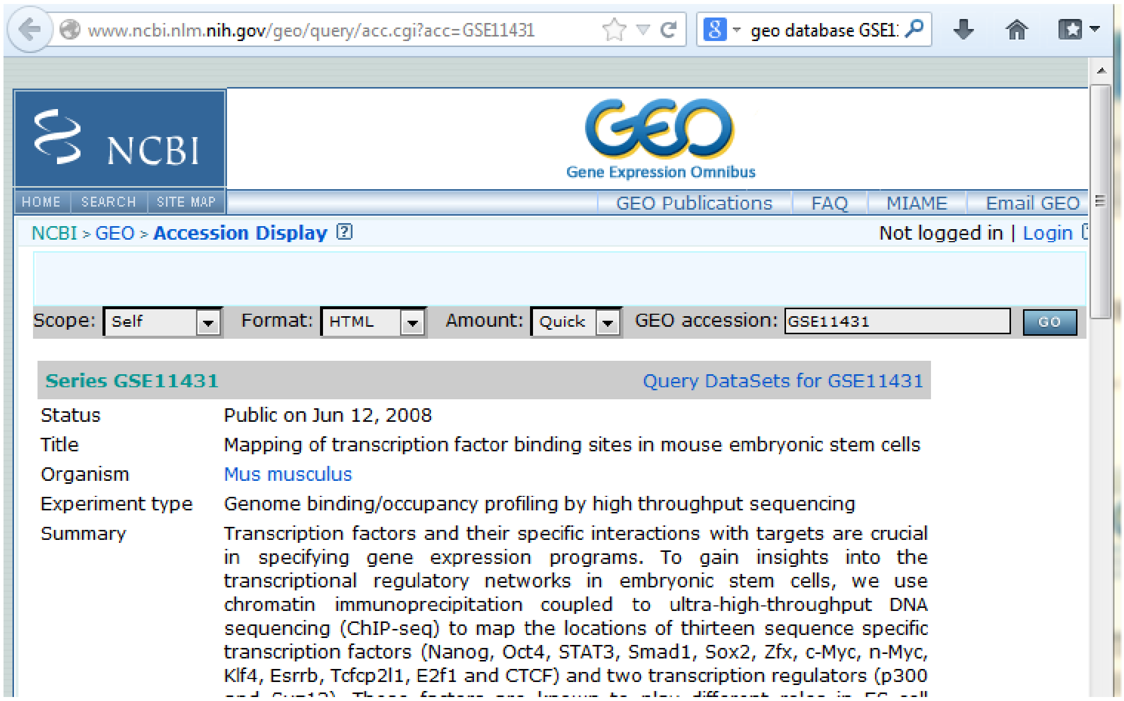
## 1.4. Goal

The focus of the current study is to provide a descriptive statistics for the combinatorial binding pattern of the various transcription factors on the genomic sites. This information can be used to lay the foundation for statistical modeling of the patterns of transcriptional factor colocalization on DNA. Such modeling treats the issue of binding of transcription factors to DNA as a Poisson point process with multiple marks, 15 total each mark for a transcription factor/coregulator. Specifically this will be a mixture model that is a mixture between a null model and point cluster model. The null model is developed which assumes that each type of points comes from a heterogeneous Poisson point processes and different binding events are independent. The null model is used to capture the marginal binding pattern of transcription factors. Then a point cluster model for a set of interacting points is assumed, which represents modules containing co-clustering of multiple transcription factor binding. The descriptive statistics will be performed with regards to the null and the point cluster model assumption and the discussion will be provided. Ultimately, the modeling process for the ChIP-seq data will provide general candidate genes meriting further wet-lab experimental investigation.

# Chapter 2. METHOD, RESULT AND DISCUSSION

## 2.1. data munging and visualization

The ChIP-seq dataset from the research by Chen et al contains the genome-wide transcription factor binding sites from the mouse embryonic stem cells (Chen et al. 2008). The data was downloaded from the GEO database with accession number GSE11431, containing a total of 15 data files representing the 15 transcription factors and coregulators (Figure 4). The url is http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11431.



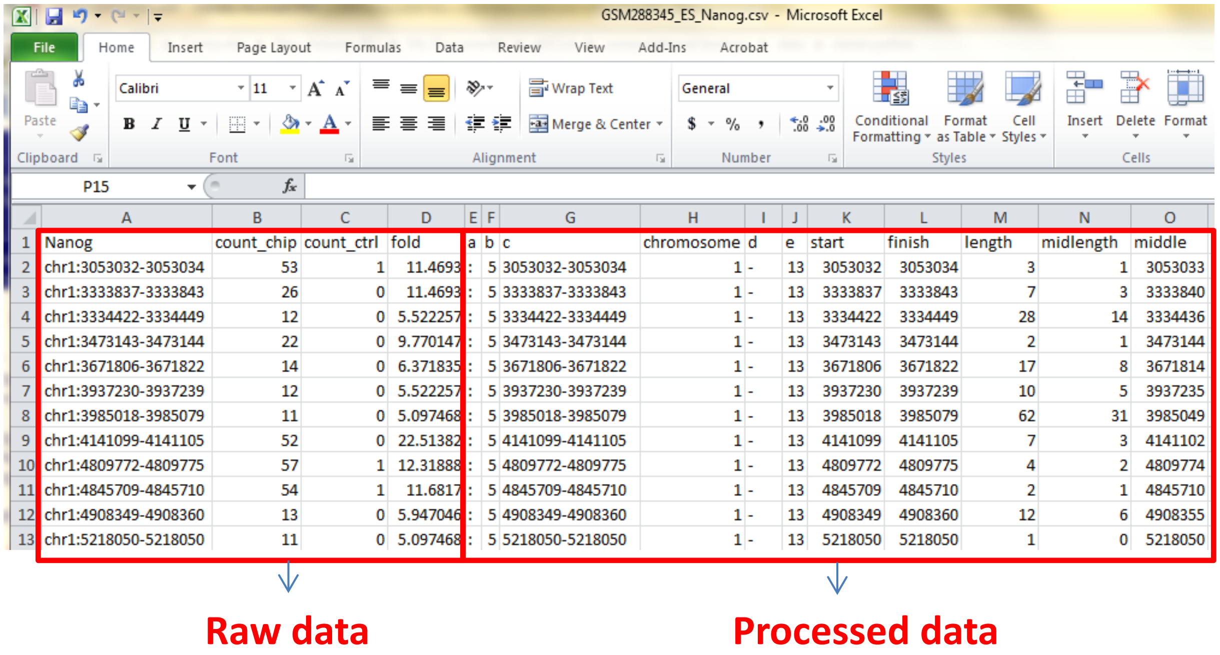
Scope: Self Format: HTML Amount: Quick GEO accession: GSE11431

Series GSE11431 Query DataSets for GSE11431

| | |
|---|---|
| Status | Public on Jun 12, 2008 |
| Title | Mapping of transcription factor binding sites in mouse embryonic stem cells |
| Organism | Mus musculus |
| Experiment type | Genome binding/occupancy profiling by high throughput sequencing |
| Summary | Transcription factors and their specific interactions with targets are crucial in specifying gene expression programs. To gain insights into the transcriptional regulatory networks in embryonic stem cells, we use chromatin immunoprecipitation coupled to ultra-high-throughput DNA sequencing (ChIP-seq) to map the locations of thirteen sequence specific transcription factors (Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1 and CTCF) and two transcription regulators (p300 and Suz12). These factors are known to play different roles in ES cell |

**Figure 4.** The data source from the Geo database.

Sample data for the genome-wide mapping of binding loci of transcription factor Nanog is shown (Figure 5). Note that the mouse cells contain 20 pairs of chromosomes, including chromosome 1 to 19 and chromosome X. Note also that the binding spans a short region and the variable “fold’ represents the area of the binding peaks and the binding strength. However, the “fold” variable was disregarded in subsequent analyses for the sake of simplicity. Additionally, the length of the binding regions, which are expressed as the difference between the start and the finish coordinates of the binding sites and have a median value of 9, was disregarded and a variable called “middle” was generated which represent the coordinate of the center of the binding site on the chromosome. From this time on, the binding was simplified as a point, which is a reasonable and necessary assumption for the future statistical modeling using a mixture model that is a mixture between the Poisson process and the point cluster model.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Nanog | count_chip | count_ctrl | fold | a | b | c | chromosome | d | e | start | finish | length | midlength | middle |
| 2 | chr1:3053032-3053034 | 53 | 1 | 11.4693 | : | 5 | 3053032-3053034 | 1 | - | 13 | 3053032 | 3053034 | 3 | 1 | 3053033 |
| 3 | chr1:3333837-3333843 | 26 | 0 | 11.4693 | : | 5 | 3333837-3333843 | 1 | - | 13 | 3333837 | 3333843 | 7 | 3 | 3333840 |
| 4 | chr1:3334422-3334449 | 12 | 0 | 5.522257 | : | 5 | 3334422-3334449 | 1 | - | 13 | 3334422 | 3334449 | 28 | 14 | 3334436 |
| 5 | chr1:3473143-3473144 | 22 | 0 | 9.770147 | : | 5 | 3473143-3473144 | 1 | - | 13 | 3473143 | 3473144 | 2 | 1 | 3473144 |
| 6 | chr1:3671806-3671822 | 14 | 0 | 6.371835 | : | 5 | 3671806-3671822 | 1 | - | 13 | 3671806 | 3671822 | 17 | 8 | 3671814 |
| 7 | chr1:3937230-3937239 | 12 | 0 | 5.522257 | : | 5 | 3937230-3937239 | 1 | - | 13 | 3937230 | 3937239 | 10 | 5 | 3937235 |
| 8 | chr1:3985018-3985079 | 11 | 0 | 5.097468 | : | 5 | 3985018-3985079 | 1 | - | 13 | 3985018 | 3985079 | 62 | 31 | 3985049 |
| 9 | chr1:4141099-4141105 | 52 | 0 | 22.51382 | : | 5 | 4141099-4141105 | 1 | - | 13 | 4141099 | 4141105 | 7 | 3 | 4141102 |
| 10 | chr1:4809772-4809775 | 57 | 1 | 12.31888 | : | 5 | 4809772-4809775 | 1 | - | 13 | 4809772 | 4809775 | 4 | 2 | 4809774 |
| 11 | chr1:4845709-4845710 | 54 | 1 | 11.6817 | : | 5 | 4845709-4845710 | 1 | - | 13 | 4845709 | 4845710 | 2 | 1 | 4845710 |
| 12 | chr1:4908349-4908360 | 13 | 0 | 5.947046 | : | 5 | 4908349-4908360 | 1 | - | 13 | 4908349 | 4908360 | 12 | 6 | 4908355 |
| 13 | chr1:5218050-5218050 | 11 | 0 | 5.097468 | : | 5 | 5218050-5218050 | 1 | - | 13 | 5218050 | 5218050 | 1 | 0 | 5218050 |

**Figure 5.** The raw data (left) and the processed data (right) shown side by side.

From the raw data indexed by transcription factors, I rearranged the data into a chromosome-specific format, such as in the table below for chromosome 1 (Table 1).

**Table 1.** Transcription factor/coregulator bindings on chromosome 1.

| chromosome | Transcription Factor | start | finish | Length (bp) | middle |
|---|---|---|---|---|---|
| 1 | Ctcf | 3002834 | 3002851 | 18 | 3002843 |
| 1 | Nanog | 3053032 | 3053034 | 3 | 3053033 |
| 1 | Sox2 | 3053046 | 3053052 | 7 | 3053049 |
| 1 | Nanog | 3333837 | 3333843 | 7 | 3333840 |
| 1 | Smad1 | 3334335 | 3334449 | 115 | 3334392 |
| 1 | Nanog | 3334422 | 3334449 | 28 | 3334436 |
| 1 | Nanog | 3473143 | 3473144 | 2 | 3473144 |
| 1 | Smad1 | 3479676 | 3479748 | 73 | 3479712 |
| 1 | 4-Oct | 3671785 | 3671827 | 43 | 3671806 |
| 1 | Nanog | 3671806 | 3671822 | 17 | 3671814 |
| 1 | Sox2 | 3904283 | 3904333 | 51 | 3904308 |
| 1 | Nanog | 3937230 | 3937239 | 10 | 3937235 |
| 1 | Sox2 | 3984948 | 3984950 | 3 | 3984949 |
| 1 | Nanog | 3985018 | 3985079 | 62 | 3985049 |
| 1 | Tcfcp2l1 | 4049605 | 4049669 | 65 | 4049637 |
| : | : | : | : | : | : |
| : | : | : | : | : | : |
| 1 | Esrrb | 196677318 | 196677323 | 6 | 196677321 |
| 1 | 4-Oct | 196701958 | 196701982 | 25 | 196701970 |
| 1 | Sox2 | 196702007 | 196702019 | 13 | 196702013 |
| 1 | Tcfcp2l1 | 196702244 | 196702246 | 3 | 196702245 |
| 1 | Ctcf | 196741335 | 196741341 | 7 | 196741338 |
| 1 | Ctcf | 196831733 | 196831794 | 62 | 196831764 |
| 1 | nMyc | 196831775 | 196831784 | 10 | 196831780 |
| 1 | Tcfcp2l1 | 196831802 | 196831813 | 12 | 196831808 |
| 1 | E2f1 | 196831821 | 196831826 | 6 | 196831824 |
| 1 | Zfx | 196831877 | 196831881 | 5 | 196831879 |
| 1 | Tcfcp2l1 | 196877403 | 196877413 | 11 | 196877408 |
| 1 | 4-Oct | 196877480 | 196877488 | 9 | 196877484 |
| 1 | Ctcf | 196933550 | 196933554 | 5 | 196933552 |
| 1 | Ctcf | 196942349 | 196942358 | 10 | 196942354 |

The clustering of the binding sites in chromosome 1 is visualized, using the "plot" function of R (Figure 6). Note the multiple clusters of transcription factor binding on chromosome 1, especially the one on the top right of the plot, with coordinate of 34084801 containing 7 transcription factors including Sta3, Tcfcp2l1, Smad1, Klf4, Oct4, E2f1, n-Myc.

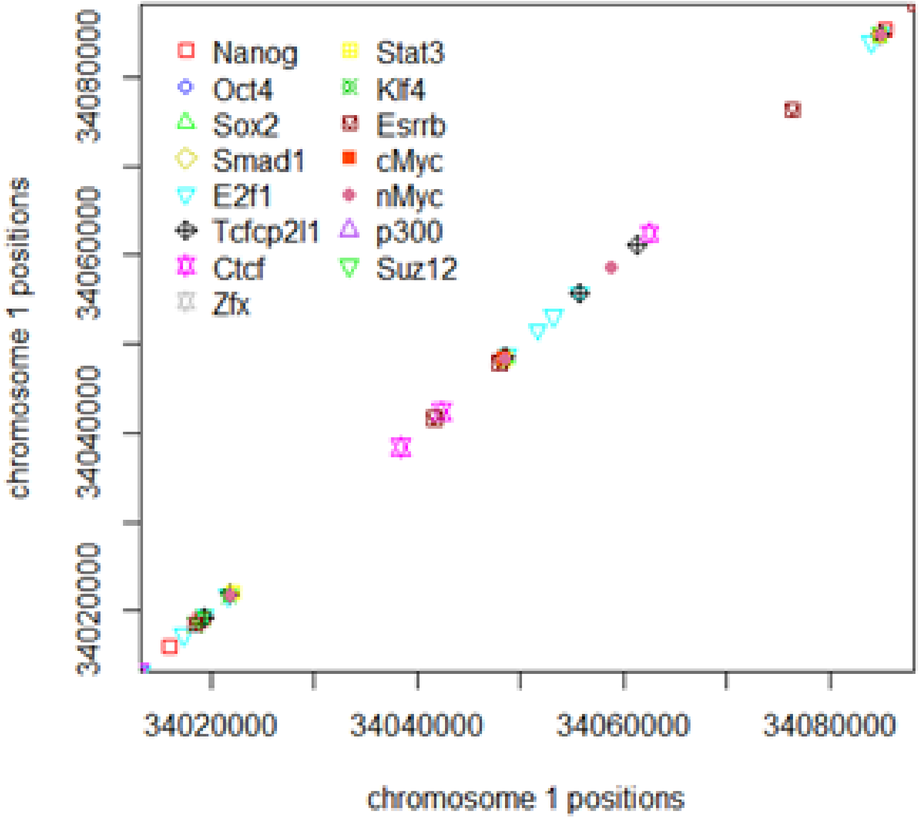
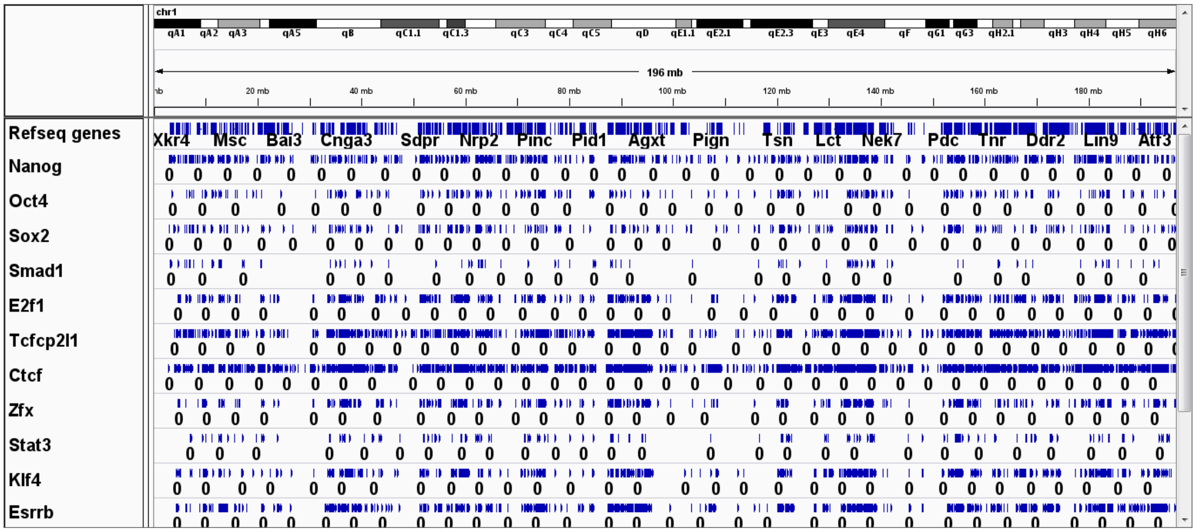


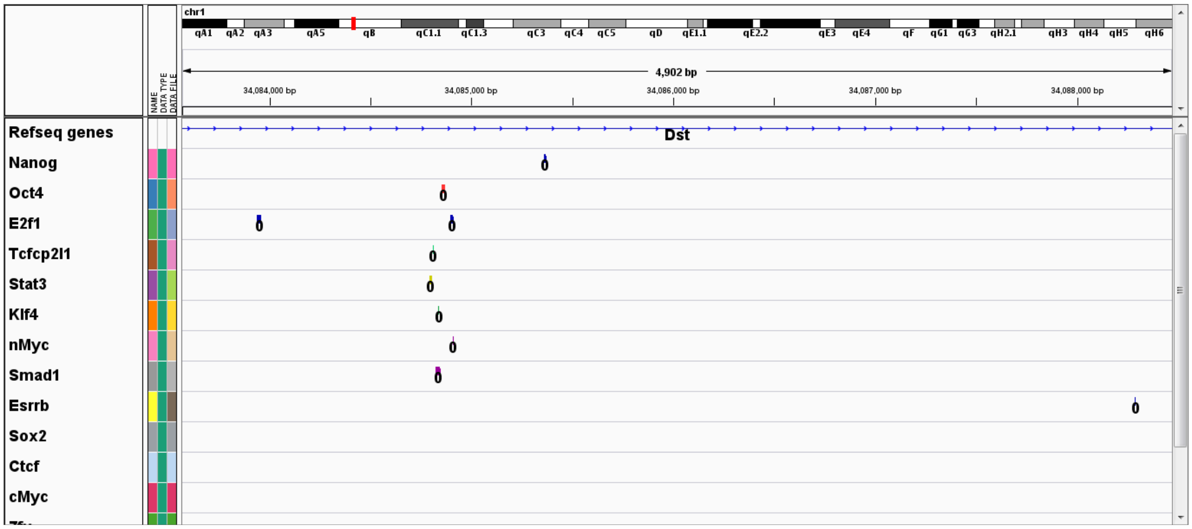
**Figure 6.** Binding of transcription factors and coregulators on chromosome 1.

However, such visualization method by generating R graphs was rather inconvenient as switching between regions requires manual input of specific region coordinates and execution of the code; thus I used another google map-like visualized tool, the genomic browser software IGV (Broad Institute, Massachusetts Institute of Technology), after converting the binding data into standard .bed file.  IGV was a much more convenient tool for visualizing the specific sites on the chromosome with the ease of zooming in and out on any target regions (Figure 7.)

A)


chr1
qA1 qA2 qA3 qA5 qB qC1.1 qC1.3 qC3 qC4 qC5 qD qE1.1 qE2.1 qE2.3 qE3 qE4 qF qG1 qG3 qH2.1 qH3 qH4 qH5 qH6
196 mb
20 mb 40 mb 60 mb 80 mb 100 mb 120 mb 140 mb 160 mb 180 mb
Refseq genes
Xkr4 Msc Bai3 Cnga3 Sdpr Nrp2 Pinc Pid1 Agxt Pign Tsn Lct Nek7 Pdc Tnr Ddr2 Lin9 Atf3
Nanog
Oct4
Sox2
Smad1
E2f1
Tcfcp2l1
Ctcf
Zfx
Stat3
Klf4
Esrrb

(B)



(C)

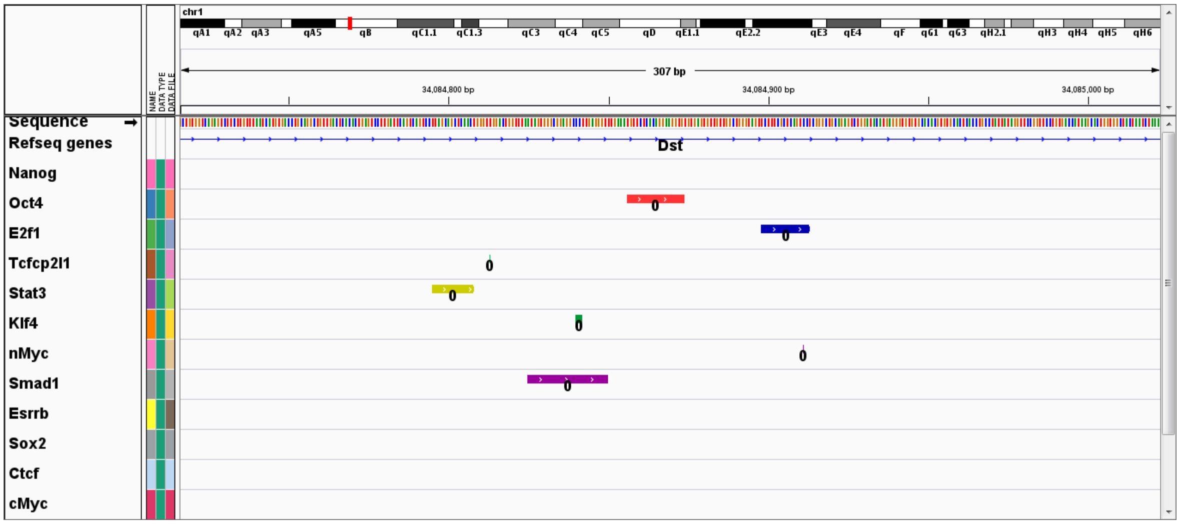


**Figure 7.** The genomic view of the transcription factor binding sites on the site 34084801 of chromosome 1 containing 7 transcription factors including Sta3, Tcfcp2l1, Smad1, Klf4, Oct4, E2f1, n-Myc , generated using the software IGV 2.2. This cluster is located within the intron region of the gene Dst. Shown are images of (A) whole chromosome 1, (B) a 4 kb region near the gene Dst, and (C) a 307 bp region around the heavily transcription factor-clustered site.

## 2.2. Test of whether the transcription factor binding sites are a Poisson process

In order to rule out the possibility that the transcription factor binding is purely due to random effect and to pave the way for the future statistical modeling using a mixture model stated above between a null model and point cluster model, the chi-square goodness of fit test for a Poisson process was conducted as described, which has been used for estimating whether the transcription factor binding sites in the *Saccharomyces cerevisiae* (baking yeast) genome constitutes a Poisson process (Wagner 1999; Sokal and Rohlf 1969). Recall that there are three equivalent ways of describing a Poisson process and one way used in the thesis research is that the interarrival (interbinding) distances of a Poisson process is exponentially distributed (Figure 6). Thus the null hypothesis is that interbinding distance is exponentially distributed.

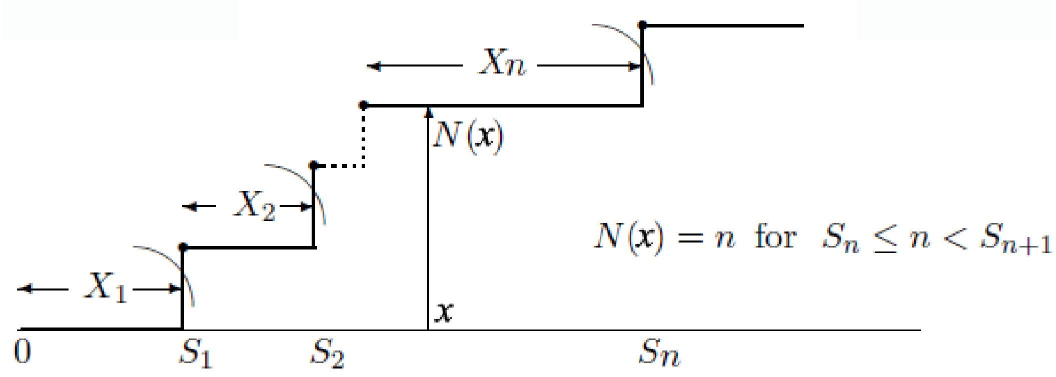


**Figure 8**. Poisson process can be described as three equivalent ways: an arrival process with arrival epoch $\{\boldsymbol{S_1}, \boldsymbol{S_2}, \ldots\}$, an interarrival (interbinding in our case) process with interval $\{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots\}$, or a counting process with the number of count up to location $\boldsymbol{x}$ $\{\boldsymbol{N(x)}; \boldsymbol{x > 0}\}$. The interarrival process is used in the thesis research.

Under the null hypothesis,

each $\mathrm{X_i}, \mathrm{i} = 1,2, \dots, \mathrm{n}, \dots$, where $\mathrm{X_i}$ is the interbinding distance between adjacent sites

with density

$$\mathbf{f_X(x) = \lambda \exp(-\lambda\, x) \;\; for\; x > 0}$$

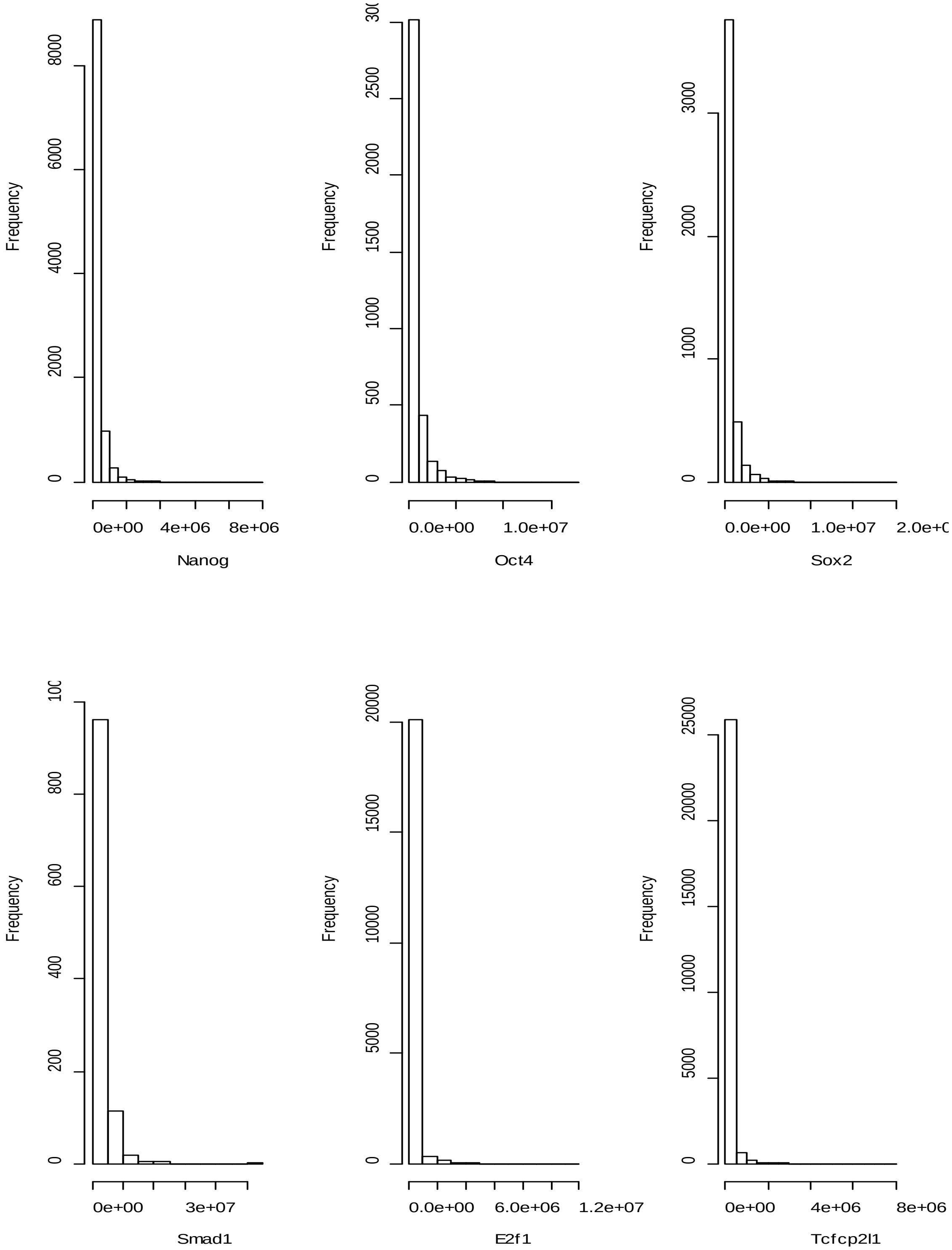
and the cumulative distribution function

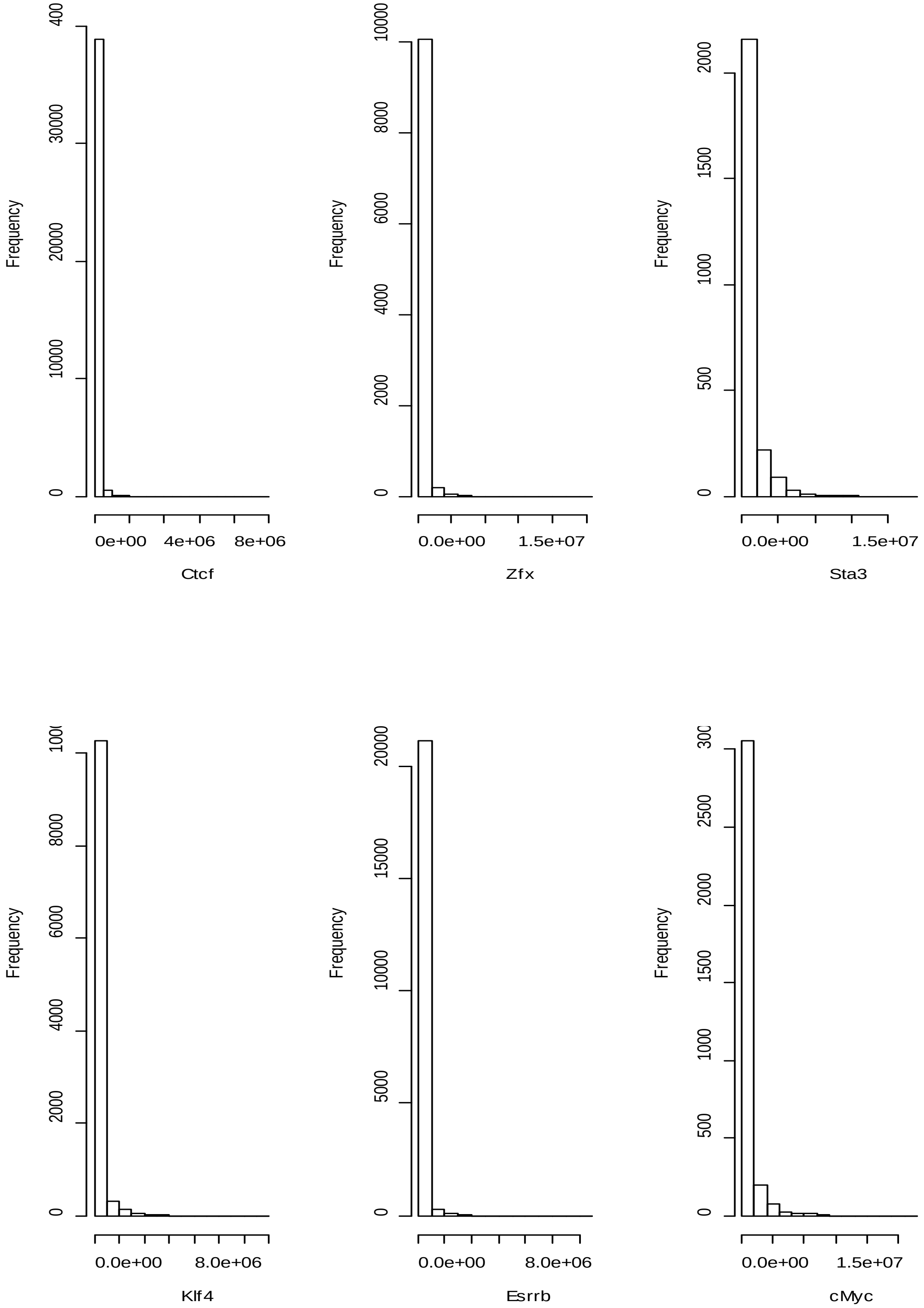
$$\mathbf{F_X(x) = 1 - \;\; \exp(-\lambda\, x) \;\; for\; x > 0}$$

where $\lambda$ is the rate

The analytical procedure is as below:

1) The data for each transcription factor, such as that in Figure 1 for the transcription factor Nanog, is converted to the interbinding distances between the adjacent binding sites. The distances for each transcription factor have the following histograms (Figure 9), that has resemblance to the density function of exponential distribution. But are they really exponentially distributed? Chi-square goodness of fit of test is to answer that.


Frequency
0
2000
4000
6000
8000
0e+00
4e+06
8e+06
Nanog
Frequency
0
500
1000
1500
2000
2500
30
0.0e+00
1.0e+07
Oct4
Frequency
0
1000
2000
3000
0.0e+00
1.0e+07
2.0e+0
Sox2
Frequency
0
200
400
600
800
10
0e+00
3e+07
Smad1
Frequency
0
5000
10000
15000
20000
0.0e+00
6.0e+06
1.2e+07
E2f1
Frequency
0
5000
10000
15000
20000
25000
0e+00
4e+06
8e+06
Tcfcp2l1


Frequency
400
30000
20000
10000
0
0e+00
4e+06
8e+06
Ctcf
Frequency
10000
8000
6000
4000
2000
0
0.0e+00
1.5e+07
Zfx
Frequency
2000
1500
1000
500
0
0.0e+00
1.5e+07
Sta3
Frequency
100
8000
6000
4000
2000
0
0.0e+00
8.0e+06
Klf4
Frequency
20000
15000
10000
5000
0
0.0e+00
8.0e+06
Esrrb
Frequency
300
2500
2000
1500
1000
500
0
0.0e+00
1.5e+07
cMyc

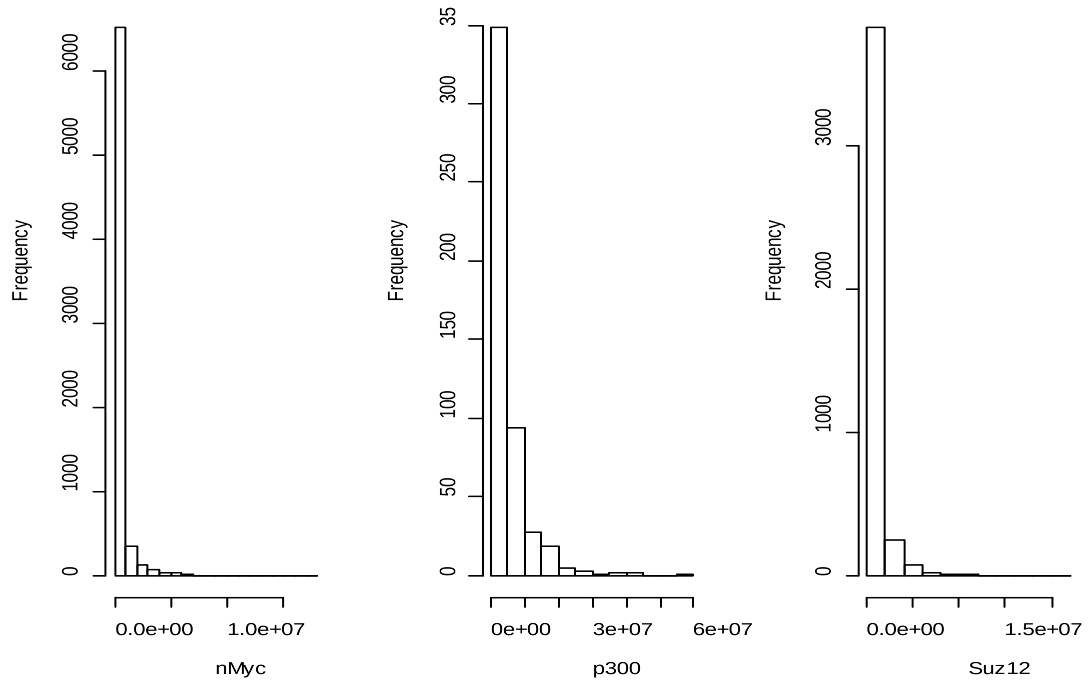


**Figure 9.** Histograms of the interval lengths for the 15 transcription factors and coregulators.

2) Then the interbinding distances are divided into various intervals according to the values of the distances. As the chi-square goodness of fit test is based on large sample approximations and is accurate only if 1) all expected values > 1 and 2) at least 80% of the expected values > 5, caution was taken to make the expected distance in each interval meet the two requirements above. Where the intervals are with expected value less than 5, they were merged with the adjacent intervals.

3) The parameter of the null exponential distribution, the rate $\boldsymbol{\lambda}$**,** is estimated based on the mean distances,

$$\boldsymbol{\lambda} = \frac{\mathbf{1}}{\textbf{the mean distance}}$$

For example, for the transcription factor Nanog, the mean distance is 246969, $\boldsymbol{\lambda} = \mathbf{1}/$ 246969**.**

The cumulative probability for Nanog in the interval (0-20000) was obtained using the cumulative distribution function of the exponential distribution

$$F_X(x) = 1 - \ exp(-\lambda\, x) = 1 - exp\left(-\frac{20000}{246956}\right) = 0.0779$$

The results for the transcription factor Nanog are shown below after dividing the interbinding distances into 11 intervals (Table 2). The Chi value is 2595. The degree of freedom (D.F.) was the total number of intervals minus 2, which equals to 9. Thus the p value (for rejecting the null hypothesis) is close to zero. Thus the hypothesis that the interbinding distances are exponentially distributed and the binding sites constitute a Poisson process is rejected for Nanog.

**Table 2.** Calculation in the chi-square goodness of fit test for Nanog.

| Nanog | Interbinding distance intervals | Interval probability | Cumulative probability | Observed | Expected | (E-O)^2/E |
|---|---|---|---|---|---|---|
| 1 | 0-20000 | 0.07779 | 0.07779 | 2071 | 803.023 | 2002.144 |
| 2 | 20000-30000 | 0.03660 | 0.11438 | 555 | 377.773 | 83.143 |
| 3 | 30,000-40,000 | 0.03514 | 0.14953 | 441 | 362.782 | 16.864 |
| 4 | 40,000-60,000 | 0.06616 | 0.21569 | 770 | 682.948 | 11.096 |
| 5 | 60,000-80,000 | 0.06101 | 0.27670 | 607 | 629.822 | 0.827 |
| 6 | 80,000-100,000 | 0.05627 | 0.33296 | 504 | 580.828 | 10.162 |
| 7 | 100,000-200,000 | 0.22210 | 0.55506 | 1753 | 2292.724 | 127.055 |
| 8 | 200,000-300,000 | 0.14815 | 0.70321 | 1041 | 1529.332 | 155.930 |
| 9 | 300,000-400,000 | 0.09882 | 0.80203 | 670 | 1020.121 | 120.167 |
| 10 | 400,000-500,000 | 0.06592 | 0.86795 | 473 | 680.459 | 63.250 |
| 11 | 500,000-Infiniy | 0.13205 | 1.00000 | 1438 | 1363.188 | 4.106 |
| Total | | | | 10323 | 10323 | Chi^2 = 2595 |

Additionally, similar analyses were performed on other transcription factors and coregulators and the results are shown below (Table 3), which is similar to that of Nanog. To ensure meeting the Chi-square goodness of test requirement, the lower bounds of the expected value E for the transcription factors are shown, which were all >5. It is concluded that the transcription factor binding on the chromosomes is likely not a Poisson process.

**Table 3**: Results of chi square goodness of fit test for all transcription factors/coregulators

| | Transcription factor(coregulator) | No. of sites | D.F. | minimal E | Chi | p value |
|---|---|---|---|---|---|---|
| 1 | Nanog | 10324 | 9 | 363 | 2595 | $<10^{-10}$ |
| 2 | Oct4 | 3742 | 9 | 53 | 1081 | $<10^{-10}$ |
| 3 | Sox2 | 4507 | 9 | 75 | 981 | $<10^{-10}$ |
| 4 | Smad1 | 1107 | 7 | 12 | 673 | $<10^{-10}$ |
| 5 | E2f1 | 20680 | 9 | 358 | 25102 | $<10^{-10}$ |
| 6 | Tcfcp2l1 | 26891 | 9 | 139 | 17070 | $<10^{-10}$ |
| 7 | Ctcf | 39590 | 9 | 17 | 36392 | $<10^{-10}$ |
| 8 | Zfx | 10319 | 9 | 363 | 9437 | $<10^{-10}$ |
| 9 | Sta3 | 2527 | 9 | 25 | 1566 | $<10^{-10}$ |
| 10 | Klf4 | 10856 | 9 | 400 | 7885 | $<10^{-10}$ |
| 11 | Esrrb | 21628 | 9 | 313 | 17510 | $<10^{-10}$ |
| 12 | cMyc | 3403 | 9 | 44 | 3715 | $<10^{-10}$ |
| 13 | nMyc | 7163 | 9 | 184 | 5923 | $<10^{-10}$ |
| 14 | p300 | 505 | 7 | 5 | 162 | $<10^{-10}$ |
| 15 | Suz12 | 4196 | 9 | 66 | 26277 | $<10^{-10}$ |

There are possibly two explanations for the results of chi square goodness of fit test: 1) the transcription factor binding sites are not a Poisson process due to the presence of many co-clustering site with great than or equal to two transcription factors/coregulators; or 2) the transcription factor binding is a Poisson process, but it is non-homogeneous with the rate parameter different at different regions of the chromosome. To discern those two possibilities, we deleted the binding sites that contains greater than or equal to two transcription factors/coregulators. The definition of a co-clustering is that if two or more transcription factor bindings sites are within 50 bp next to each other, they are considered to be co-clustered (For further discussion, please read section 2.2). The new results are shown (Table 4) and are similar to that in Table 3, indicating that the null hypothesis is still to be rejected. Such results may suggest that the second possibility above is more likely, as removal of co-clustering does not make a difference. Such results provide guidance for the choice of models in future modeling study of transcription factor binding sites.

**Table 4.** Results of chi square goodness of fit test for binding sites after deleting the transcription factor co-clustering

| | Transcription factor | No. of sites | d.f. | minimal E | Chi | p value |
|---|---|---|---|---|---|---|
| 1 | Nanog | 6543 | 9 | 154 | 849 | $<10^{-10}$ |
| 2 | 4-Oct | 1272 | 9 | 6 | 161 | $<10^{-10}$ |
| 3 | Sox2 | 1435 | 9 | 8 | 156 | $<10^{-10}$ |
| 4 | Smad1 | 283 | 7 | 2 | 58 | $4.2\times10^{-10}$ |
| 5 | E2f1 | 15357 | 9 | 621 | 20597 | $<10^{-10}$ |
| 6 | Tcfcp2l1 | 21615 | 9 | 314 | 10470 | $<10^{-10}$ |
| 7 | Ctcf | 37443 | 9 | 25 | 27144 | $<10^{-10}$ |
| 8 | Zfx | 7374 | 9 | 194 | 6649 | $<10^{-10}$ |
| 9 | Sta3 | 1220 | 9 | 6 | 287 | $<10^{-10}$ |
| 10 | Klf4 | 6414 | 9 | 149 | 4059 | $<10^{-10}$ |
| 11 | Esrrb | 16964 | 9 | 577 | 12912 | $<10^{-10}$ |
| 12 | cMyc | 1093 | 9 | 5 | 682 | $<10^{-10}$ |
| 13 | nMyc | 3059 | 9 | 36 | 1756 | $<10^{-10}$ |
| 14 | p300 | 259 | 7 | 2 | 115 | $<10^{-10}$ |
| 15 | Suz12 | 3968 | 9 | 59 | 26072 | $<10^{-10}$ |

2.3. Patterns of multiple transcription factor binding sites based on correlation analysis.

It was noted from the data that many transcription factor binding sites are clustered (Figure 7). In order to model the clustering of transcription factor/coregulator binding sites, it is necessary to delve into the various patterns of multiple transcription factor co-clustering. First, I defined a cluster as follows: two transcription binding sites that are within 50 bp to each other are considered as within a cluster. Based on this definition, I next introduced an indicator variable for whether there is a cluster:

$$r = 1(x_j - x_{j-1} \leq \beta), where\ \beta = 50\ bp$$

$$j = 2, 3, \dots, n, \dots$$

i.e.

$$r = \begin{cases} 1 & x_j - x_{j-1} \leq \beta \\ 0 & otherwise \end{cases}$$

Additionally, another variable w is defined, which is the index of the points in a clustered site.

For all points in a clustered site, $r$ = 1 and $w \in \{1, 2, 3, \dots n, \dots\}$

For the initial point in a cluster, $w = 1$

For the second point in a cluster, $w = 1$

:
:

For the n-th point next to the initial point, $w = n$

For all points not in a cluster, $r$ = 0, $w = 0$

The chromosome 1 data in Table 1 is thus transformed after calculating all the $r$ and $w$ values. The same transformation is repeated to other chromosomes, 1 to 19 and X. These transformed chromosome-specific datasets were stacked together (Table 5). The maximal value for w is 10. That means that there are co-clustered sites with 10 transcription factor/coregulators. In total, there are 17434 clusters and the largest cluster contains 10 transcription factor/cofactors. In the seminal paper by Chen et al, a distance of 100 bp or shorter was used (Chen et al. 2008). Using a distance of 100 bp or shorter resulted in more clusters found, 22462 in total, and the largest cluster contains 12 transcription factors/coregulators. Because one major purpose of statistical analysis is to generate prediction/hypothesis for wet-lab biological research to test, I believe that a more stringent definition of 50 bp or shorter would be more beneficial and was adopted for subsequent analyses.

**Table 5.** Binding sites of transcription factors on chromosome 1 to 19 and X

| | chromosome | TF | middle | r | w |
|---|---|---|---|---|---|
| 1 | 1 | Ctcf | 3002843 | 0 | 0 |
| 2 | 1 | Nanog | 3053033 | 1 | 1 |
| 3 | 1 | Sox2 | 3053049 | 1 | 2 |
| 4 | 1 | Nanog | 3333840 | 0 | 0 |
| 5 | 1 | Smad1 | 3334392 | 1 | 1 |
| 6 | 1 | Nanog | 3334436 | 1 | 2 |
| 7 | 1 | Nanog | 3473144 | 0 | 0 |
| 8 | 1 | Smad1 | 3479712 | 0 | 0 |
| 9 | 1 | Oct4 | 3671806 | 1 | 1 |
| : | : | : | : | : | : |
| : | : | : | : | : | : |
| 42881 | 5 | Zfx | 3225503 | 1 | 1 |
| 42882 | 5 | Tcfcp2l1 | 3225522 | 1 | 2 |
| 42883 | 5 | Ctcf | 3289269 | 0 | 0 |
| 42884 | 5 | Nanog | 3299820 | 0 | 0 |
| 42885 | 5 | E2f1 | 3299942 | 0 | 0 |
| : | : | : | : | : | : |
| : | : | : | : | : | : |
| 167714 | X | Klf4 | 165341139 | 1 | 1 |
| 167715 | X | Suz12 | 165341166 | 1 | 2 |
| 167716 | X | Tcfcp2l1 | 165341230 | 1 | 1 |
| 167717 | X | nMyc | 165341253 | 1 | 2 |
| 167718 | X | Zfx | 165342022 | 0 | 0 |
| 167719 | X | Esrrb | 165342198 | 0 | 0 |
| 167720 | X | nMyc | 165349433 | 1 | 1 |
| 167721 | X | Esrrb | 165349436 | 1 | 2 |
| 167722 | X | Klf4 | 165349436 | 1 | 3 |
| 167723 | X | Tcfcp2l1 | 165349439 | 1 | 4 |

The data in Table 5 was further transformed to demonstrate the transcription factor composition of each clustered site (Table 6). A value of 1 for transcription factors indicates the presence and 0 indicates the absence of transcription factors. The column with the name “sites” are the coordinate of the left-most transcription factor-binding site in a cluster. Then, the data on Table 5 was processed using correlation analysis to demonstrate the trend of co-localization of various transcription factors/coregulators in the clustered sites. For example, if transcription factors A and B are together in all the clustered sites, their correlation is 1. Correlation matrices were generated for the clustered site with a definition of 50 bp apart ($\beta$=50 bp. Table 7). A heat map was generated based on the correlation matrices, with yellow colors indicating high likelihood of the transcription factors/coregulators to be co-localized in the clustered sites and red low likelihood (Figure 10). The heat map showed the presence of two clustered groups. One includes Nanog, Oct4 and Sox2 (blue box of Figure 10); the other includes n-Myc, c-Myc (green box of Figure 10). In the paper by Chen et al (Chen et al. 2008) using 100 bp as the definition of clustered sites, one similar group was identified, which included Nanog, Oct4, Sox2, Smad1, and STAT3; and another group induced n-Myc, c-Myc, E2f1, and Zfx. Our results are consistent with the findings by Chen et al. And it is likely due to the higher stringency of 50 bp in our definition that our clustered groups have fewer transcription factors than those by Chen et al.

**Table 6.** Transcription factors composition of the clustered sites on the mouse chromosomes.

| chromosome | sites | Nanog | Oct4 | Sox2 | Smad1 | E2f1 | Tcfcp2l1 | Ctcf | Zfx | Sta3 | Klf4 | Esrrb | cMyc | nMyc | p300 | Suz12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3053033 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 3334392 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 3671806 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 3984949 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 4141031 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 4406947 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| 5 | 3225503 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 3339191 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 3392825 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 3549717 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 3602116 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| X | 165327153 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| X | 165330874 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| X | 165334408 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| X | 165341139 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| X | 165349433 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |

**Table 7.** Correlation of transcription factor co-occurrence in the clusters (defined as 50 bp apart).

| | Nanog | Oct4 | Sox2 | Smad1 | E2f1 | Tcfcp2l1 | Ctcf | Zfx | Sta3 | Klf4 | Esrrb | cMyc | nMyc | p300 | Suz12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nanog | 1 | 0.226 | 0.467 | 0.246 | -0.265 | -0.080 | -0.176 | -0.212 | 0.031 | -0.050 | -0.039 | -0.180 | -0.241 | 0.130 | -0.057 |
| Oct4 | 0.226 | 1 | 0.308 | 0.168 | -0.131 | -0.076 | -0.084 | -0.118 | 0.026 | -0.057 | -0.084 | -0.093 | -0.111 | 0.114 | -0.034 |
| Sox2 | 0.467 | 0.308 | 1 | 0.201 | -0.229 | -0.086 | -0.131 | -0.184 | 0.011 | -0.088 | -0.093 | -0.162 | -0.216 | 0.134 | -0.047 |
| Smad1 | 0.246 | 0.168 | 0.201 | 1 | -0.115 | -0.019 | -0.080 | -0.095 | 0.095 | 0.022 | 0.028 | -0.081 | -0.101 | 0.211 | -0.026 |
| E2f1 | -0.265 | -0.131 | -0.229 | -0.115 | 1 | -0.134 | -0.146 | 0.187 | -0.084 | -0.067 | -0.210 | 0.103 | 0.156 | -0.062 | -0.053 |
| Tcfcp2l1 | -0.080 | -0.076 | -0.086 | -0.019 | -0.134 | 1 | -0.038 | -0.130 | 0.007 | -0.077 | 0.095 | -0.167 | -0.199 | 0.006 | 0.001 |
| Ctcf | -0.176 | -0.084 | -0.131 | -0.080 | -0.146 | -0.038 | 1 | -0.076 | -0.061 | -0.026 | -0.045 | -0.088 | -0.033 | -0.043 | -0.011 |
| Zfx | -0.212 | -0.118 | -0.184 | -0.095 | 0.187 | -0.130 | -0.076 | 1 | -0.068 | -0.099 | -0.113 | 0.041 | 0.040 | -0.041 | -0.009 |
| Sta3 | 0.031 | 0.026 | 0.011 | 0.095 | -0.084 | 0.007 | -0.061 | -0.068 | 1 | 0.039 | 0.031 | -0.065 | -0.070 | 0.104 | -0.027 |
| Klf4 | -0.050 | -0.057 | -0.088 | 0.022 | -0.067 | -0.077 | -0.026 | -0.099 | 0.039 | 1 | -0.012 | -0.105 | -0.082 | 0.033 | -0.037 |
| Esrrb | -0.039 | -0.084 | -0.093 | 0.028 | -0.210 | 0.095 | -0.045 | -0.113 | 0.031 | -0.012 | 1 | -0.163 | -0.213 | 0.039 | -0.015 |
| cMyc | -0.180 | -0.093 | -0.162 | -0.081 | 0.103 | -0.167 | -0.088 | 0.041 | -0.065 | -0.105 | -0.163 | 1 | 0.459 | -0.030 | -0.035 |
| nMyc | -0.241 | -0.111 | -0.216 | -0.101 | 0.156 | -0.199 | -0.033 | 0.040 | -0.070 | -0.082 | -0.213 | 0.459 | 1 | -0.048 | -0.026 |
| p300 | 0.130 | 0.114 | 0.134 | 0.211 | -0.062 | 0.006 | -0.043 | -0.041 | 0.104 | 0.033 | 0.039 | -0.030 | -0.048 | 1 | -0.005 |
| Suz12 | -0.057 | -0.034 | -0.047 | -0.026 | -0.053 | 0.001 | -0.011 | -0.009 | -0.027 | -0.037 | -0.015 | -0.035 | -0.026 | -0.005 | 1 |

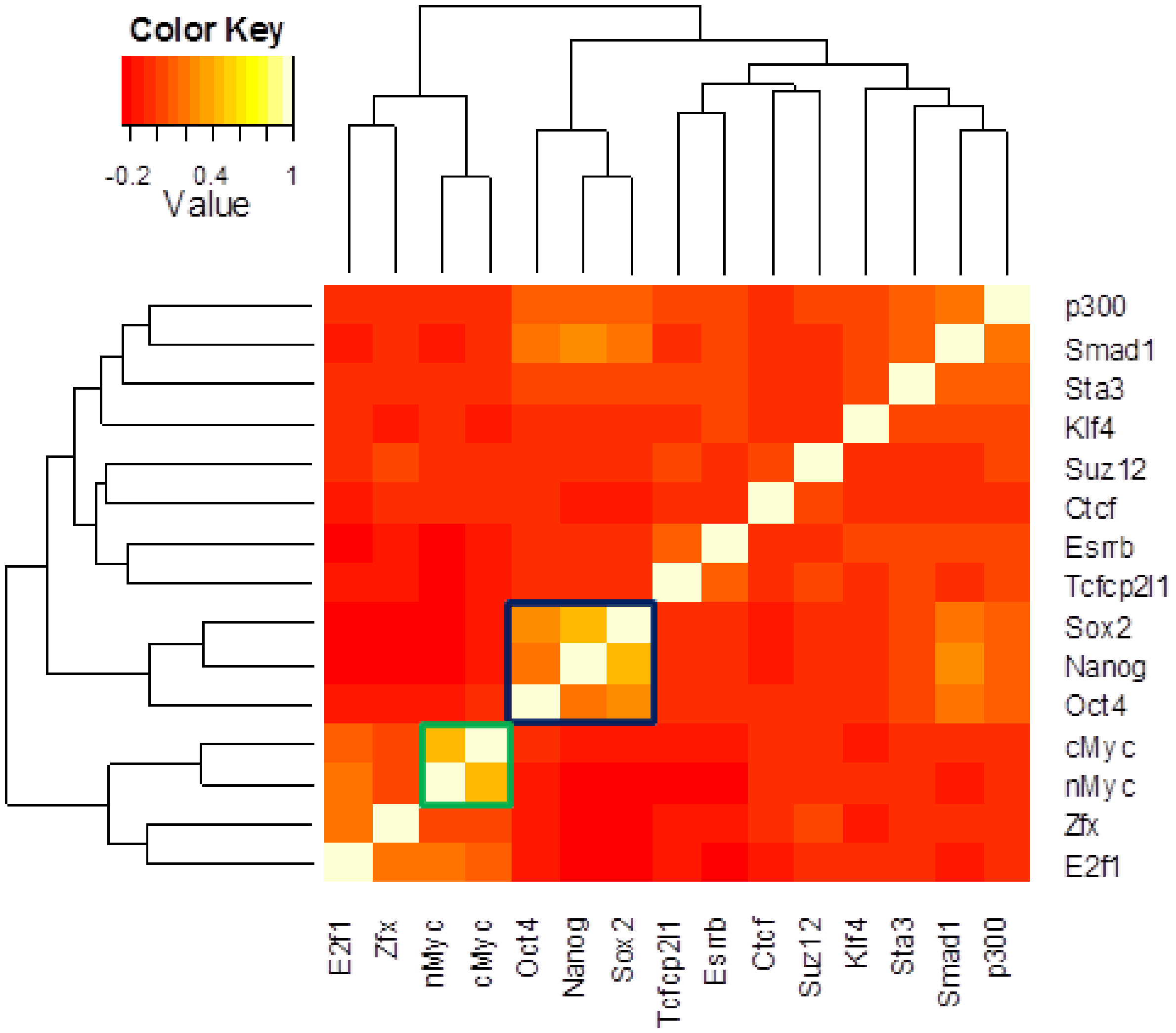


**Figure 10.** Co-occurrence of transcription factors within clustered sites (β=50 bp). Color reflects the frequency of co-localization (yellow means more likely to be co-localized, while red means less). The two patterns of transcription factor co-occurrence are marked with a blue and a green box.

2.4. Enumerations of the patterns in the multiple transcription factor binding sites.

The report by Chen et al characterized extensively the clustered sites containing 4 or more transcription factors/coregulators. However, those containing 3 transcription factors have not been explored and are addressed in the thesis research. Using my more stringent definition of clustered sites based on a distance of 50 bp or shorter, there are a total of 1856 clustered sites containing 4 or more transcription factor/coregulator and a total of 3353 containing 3 transcription factors/coregulators (Table 8).

**Table 8.** Types of clustered sites

| Number of transcription factors in a clustered site | Total clusters | Total |
|---|---|---|
| 2 | 12225 | 12225 |
| 3 | 3353 | 3353 |
| 4 | 1131 | 1856 |
| 5 | 419 | |
| 6 | 187 | |
| 7 | 73 | |
| 8 | 36 | |
| 9 | 9 | |
| 10 | 1 | |

Among the 3353 clusters with three transcription factors, there are 251 unique types. The top 20 types are listed (Table 9). Consistent with the heatmap result (Figure 10), the no. 1 ranked cluster type contains n-Myc, c-Myc, and E2f1. The no. 2 ranked cluster type contains Nanog, Oct4, and Sox2. The no. 3 ranked cluster type contains n-Myc, Zfx, and E2f1. This information will be directly relevant for the mixture modeling to find the point clusters.

Venn diagrams were also generated to show the presence of the two groups of transcription factors (group 1: Nanog, Oct4, and Sox2; group 1: n-Myc and c-Myc) in various types of clustered sites. Additionally, the same information is also expressed in pie chart (Figure 11). Additionally, for the clustered sites with 3 transcription factors, given the presence of one specific transcription factor/coregulator, the relative frequencies of the presence and absence of the two groups of transcription factors were shown in the pie chart (Figure 12). The results are showing that the transcription factor Smad1, Tcfcp1l1, Stat3, Klf4, Esrrb and the coregulator p300 are preferentially co-localized with Nanog, Oct4, Sox2, while E2f, Zfx are preferentially colocalized with n-Myc and c-Myc.

**Table 9.** Top 20 patterns of transcription factor clustering (The top 3 are highlighted). Counts mean the number of occurrence of the combination of the transcription factors. Under each transcription factor/coregulator, “1”means binding, “0” no binding.

| counts | Nanog | Oct4 | Sox2 | Smad1 | E2f1 | Tcfcp2l1 | Ctcf | Zfx | Sta3 | Klf4 | Esrrb | cMyc | nMyc | p300 | Suz12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 305 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 278 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 117 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 98 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 93 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 91 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 74 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 73 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 59 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 59 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 57 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 54 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 53 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 46 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 40 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 40 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

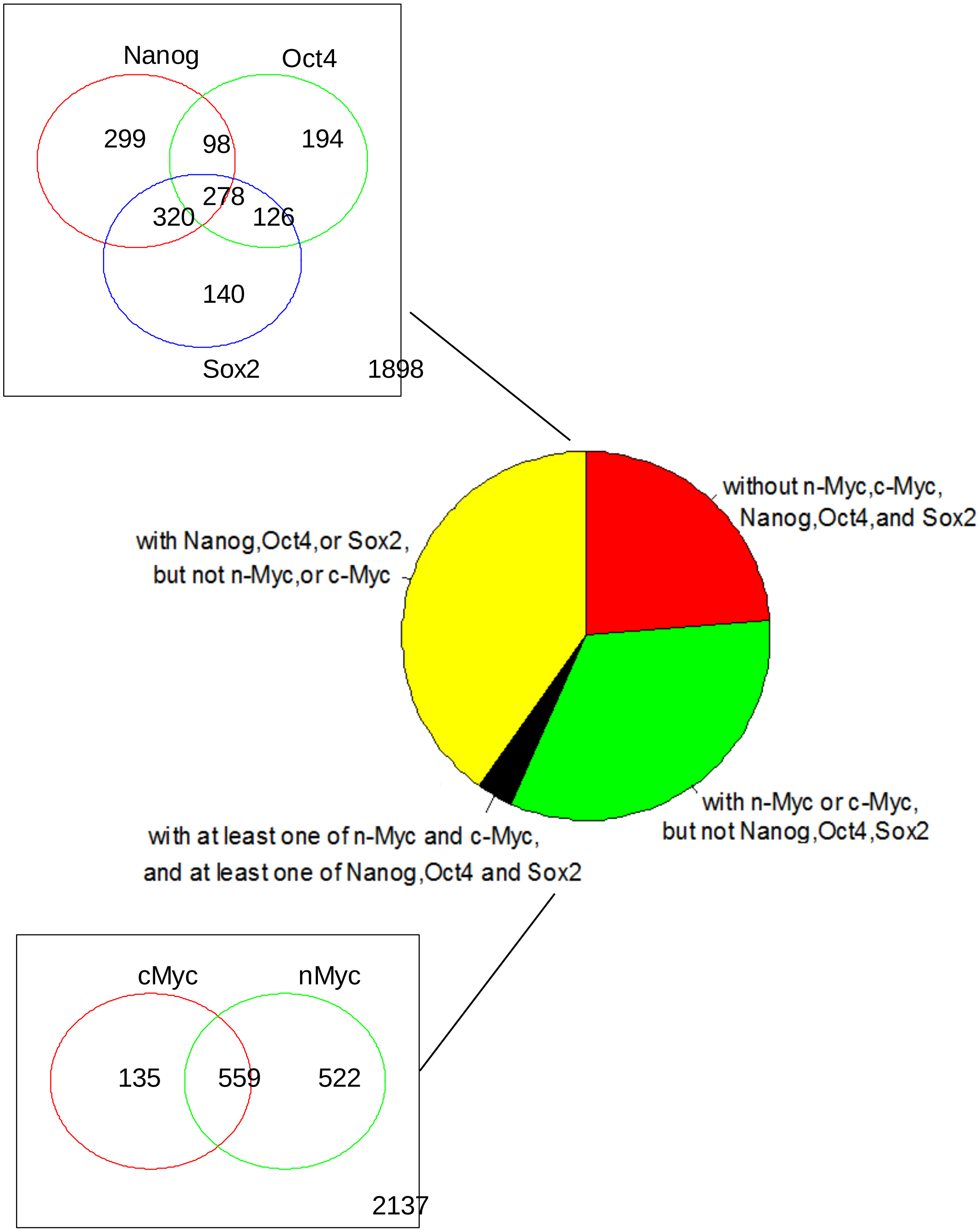
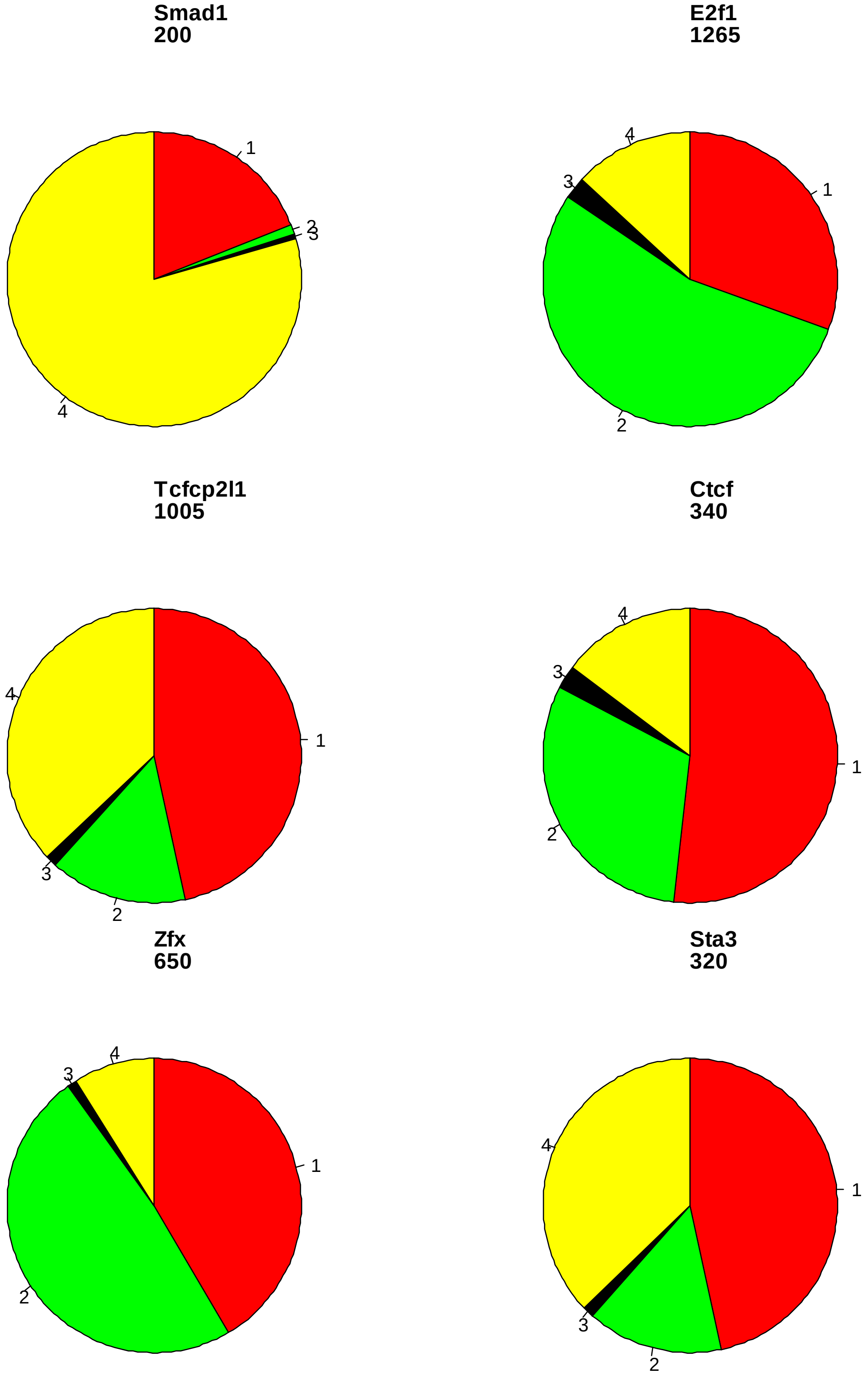
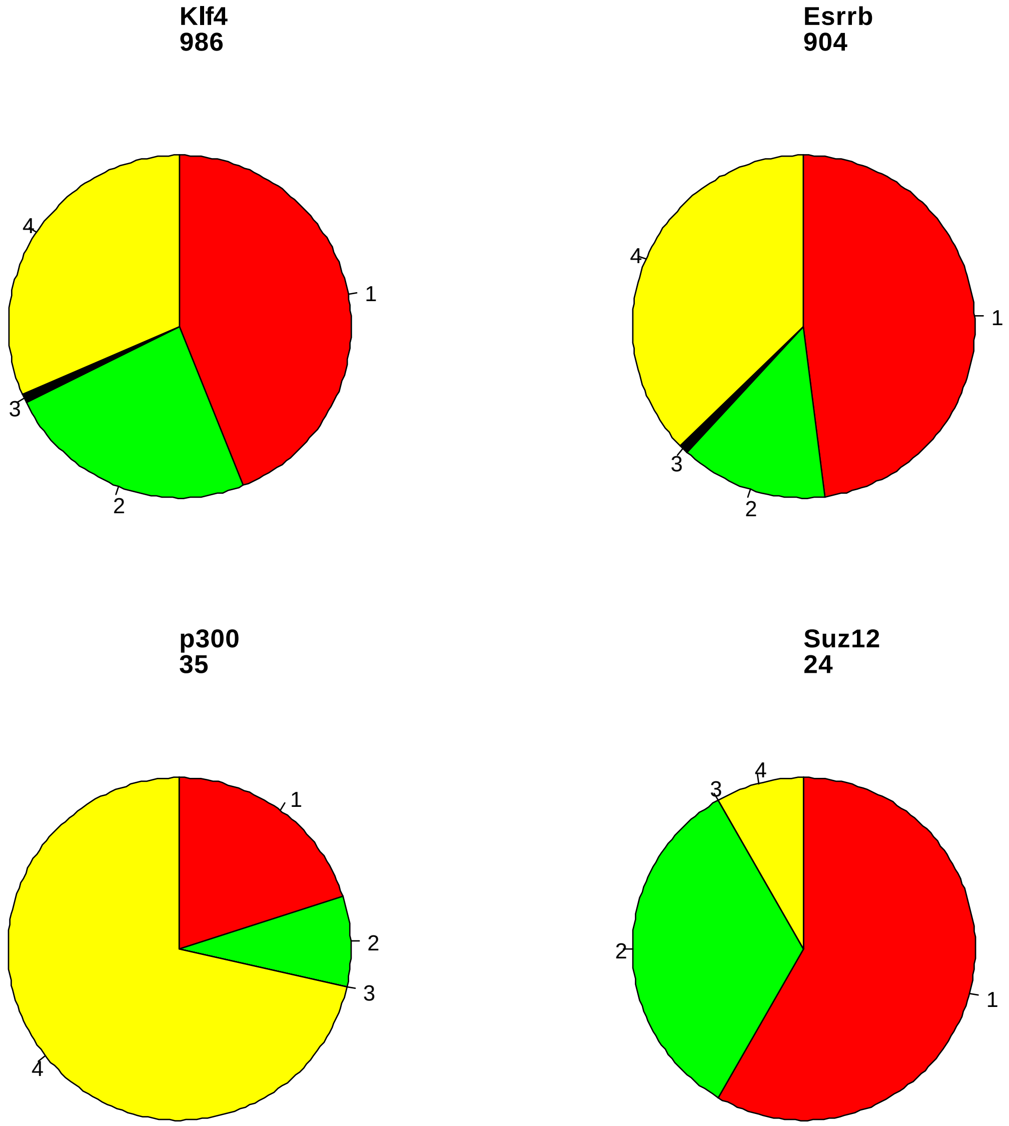


**Figure 11.** Venn diagrams for the composition of the two major co-clustering groups. The relative frequency of their occurrence in the mouse genome was shown in the pie chart.


Smad1
200
1
2
3
4
E2f1
1265
1
2
3
4
Tcfcp2l1
1005
1
2
3
4
Ctcf
340
1
2
3
4
Zfx
650
1
2
3
4
Sta3
320
1
2
3
4



1. without n-Myc,c-Myc,Nanog,Oct4,and Sox2
2. with n-Myc or c-Myc, but not Nanog,Oct4,Sox2,
3. with at least one of n-Myc and c-Myc, and at least one of Nanog,Oct4 and Sox2
4. "with Nanog,Oct4,or Sox2, but not n-Myc, or c-Myc

**Figure 12.** The relative frequency of the occurrence of the transcription factors/coregulators in the clustered sites. The names of the transcription factors/coregulators and the total number of sites containing the transcription factor/coregulators were indicated above the pie charts.

# Chapter 3. CONCLUSION

Analysis of the ChIP-seq data results in the following findings

1. The hypothesis that transcription factor/coregulator binding sites in mouse embryonic stem cells in the genome wide constitute a Poisson process is rejected. This is likely due to the many co-clustered sites or the binding being a non-homogeneous Poisson process.
2. Our correlation analysis indicated the presence of two groups of transcription factors that are preferentially colocalized. The first group contains the transcription factors Nanog, Oct4, and Sox; the second contains n-Myc and c-Myc.
3. Extensive study of the clustered binding sites containing three transcription factors was performed. It is found that there are a total of 3353 such sites. And the top two frequently co-localized groups are 1) Nanog, Oct4 and Sox 2) n-Myc, c-Myc, and E2f. The transcription factors Smad1, Tcfcp1l1, Stat3, Klf4, Esrrb and the coregulator p300 are preferentially co-localized with Nanog, Oct4, Sox2, while E2f, Zfx are preferentially colocalized with n-Myc and c-Myc. This paved the way for a statistical modeling using a mixture model between a Poisson process and a point cluster model.

# REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129 (4):823-837.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W. W., Jiang, J. M., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y. J., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., and Ng, H. H. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133 (6):1106-1117.

Davies, B. W., Bogard, R. W., and Mekalanos, J. J. 2011. Mapping the regulon of Vibrio cholerae ferric uptake regulator expands its known network of gene regulation. *Proceedings of the National Academy of Sciences* 108 (30):12467-12472.

Hobert, O. 2008. Gene Regulation by Transcription Factors and MicroRNAs. *Science* 319 (5871):1785-1786.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 316 (5830):1497-1502.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. 2008. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews* 72 (4):557-578.

McKenna, N. J., and O'Malley, B. W. 2002. Combinatorial Control of Gene Expression by Nuclear Receptors and Coregulators. *Cell* 108 (4):465-474.

Reményi, A., Scholer, H., and Wilmanns, M. 2004. Combinatorial control of gene expression. *Nature structural & molecular biology* 11 (9):812-815.

Sokal, R. R., and Rohlf, F. J. 1969. Biometry; the principles and practice of statistics in biological research, *A Series of books in biology*. San Francisco,: W. H. Freeman.

Takahashi, K., and Yamanaka, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell* 126 (4):663-676.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., and Holt, R. A. 2001. The sequence of the human genome. *science* 291 (5507):1304-1351.

Wagner, A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 15 (10):776-784.

Wilbanks, E. G., and Facciotti, M. T. 2010. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *Plos One* 5 (7).

Young, Richard A. 2011. Control of the Embryonic Stem Cell State. *Cell* 144 (6):940-954.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., and Stewart, R. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318 (5858):1917-1920.

Zhou, Q., Chipperfield, H., Melton, D. A., and Wong, W. H. 2007. A gene regulatory network in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences* 104 (42):16438-16443.