Computational analysis of ChIP-seq data and its application to the reconstruction of transcriptional regulatory networks

Xu, Han

2011

Xu, H. (2011). Computational analysis of ChIP-seq data and its application to the reconstruction of transcriptional regulatory networks. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/45765>

<https://doi.org/10.32657/10356/45765>

**Computational Analysis of ChIP-seq Data
and Its Application to the Reconstruction of
Transcriptional Regulatory Networks**

Xu Han

School of Computer Engineering

2011

**Computational Analysis of ChIP-seq Data
and Its Application to the Reconstruction of
Transcriptional Regulatory Networks**

Xu Han

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

June 12th, 2011

Acknowledgements

Firstly I would like to express my gratitude to my advisor, Prof. Lin Feng, for his guidance on my study and his valuable advice on my research work. As a part-time PhD student, I have encountered many difficulties in balancing my job and study. At these moments, Prof. Lin always encouraged me to keep pursuing my goal and showed great patience in tolerating my delay in making progress.

Secondly, I would like to thank Prof. Jagath C. R., Prof. Brusic V., Prof. Kolatkar P. R., Prof. Tan Patrick, Prof. Bertil Schmidt, Prof. Limsoon Wong, and Prof. Li Jinyan. I learnt a lot from their courses, which enlightened me from time to time in solving problems in my research work.

Thirdly, I would like to thank my colleagues in the Genome Institute of Singapore: Prof. Sung W. K, who's my supervisor in GIS and an excellent group leader; Dr. Wei C. L., Dr. Ruan Y. J., Dr. Chiu K. P., and Prof. Ng H. K., who are great biologists and generated beautiful datasets for my analysis.

Finally, special thanks to my family: my parents, my parents-in-law, my wife and two children.

Xu Han, SCE

June, 2011

Table of Contents

Table of Contents	5
Abstract	8
Author's Publications Related to Thesis	9
Glossary	10
Chapter 1: Introduction	13
1.1. Development of the ChIP-seq Technology	13
1.2. ChIP-seq Analysis Tools in Literature	18
1.3. Reconstruction of Transcriptional Regulatory Networks	23
1.4. Objectives of the Thesis	25
1.5. Major Contributions	25
1.6. Organization of the Thesis	26
Chapter 2: Single Library Analysis	28
2.1. Peak-finding for ChIP-PET Analysis	28
2.2. ChIP-seq Profiling with Virtual Fragment	32
2.3. Peak-calling	36
2.4. Application to ChIP-PET Histone Modification Dataset.....	38
2.5. Application to ChIP-seq Transcription Factor Dataset	41

Chapter 3: Comparing ChIP Library with Control	44
3.1. Intrinsic Bias of ChIP-seq	44
3.2. The Signal-noise Model of ChIP-seq	48
3.3. Noise Rate Estimation	52
3.4. FDR Estimation by Library Swapping	54
3.5. The CCAT Framework	56
3.6. Spike-in Simulation	60
3.7. Application to Histone Modifications	67
3.8. Noise Rates as a Measurement of Data Quality	73
Chapter 4: Identifying Differential Histone Modification Sites	75
4.1. Differential Histone Modification Sites	75
4.2. Determining Putative Histone Modification Sites	77
4.3. Quantitative Comparison of Intensity by Fold-change	79
4.4. An HMM-based Approach for Identifying DHMSs	83
4.5. Evaluation with H3K27me3 Data	86
4.6. Application to H3K4me3 and H3K36me3 Data	92
Chapter 5: Reconstruction of Transcriptional Regulatory Network	96
5.1. <i>De Novo</i> Motif Discovery	96

5.2. Co-localization of Multiple Transcription Factors	100
5.3. Correlation of Binding Events and Gene Expression	103
5.4. The Core Transcriptional Regulatory Network in mESC	110
Chapter 6: Discussion and Future Work	115
6.1. Summary of Thesis Work	115
6.2. Limitations	119
6.3. Future Work	122
References	126
APPENDIX I	136
APPENDIX II	139
APPENDIX III	141
APPENDIX IV	145

Abstract

On benefit of the rapid growth of ultra-high-throughput sequencing technologies in recent years, ChIP-seq (Chromatin Immuno-Precipitation Sequencing) has become the main stream for the genome-wide study of protein-DNA interactions and histone modifications. Large amount of ChIP-seq datasets have been generated and published. The analysis of ChIP-seq data posed new challenges to the bioinformatics community.

In this thesis work, we proposed three computational techniques for the analysis of ChIP-seq data, in respect to different experimental design. Firstly, we developed Peak-finder for the prediction of transcription factor binding sites from a single ChIP library; secondly, we proposed a signal-noise model of ChIP-seq, from which we derived a general-purpose framework, CCAT (Control based ChIP-seq Analysis Tool), for the ChIP-seq applications with negative controls; thirdly, we introduced an HMM (Hidden Markov Model) approach, named ChIPDiff, to the comparative analysis of two ChIP-seq libraries that are associated with different cell-types or treatments.

Next we addressed the problem of the reconstruction of transcriptional regulatory networks, which depict the relationship between transcription factors and their target genes. Based on the prediction of ChIP-enriched genomic sites, we proposed a probabilistic method to link the transcription factor binding sites to the putative target genes. We further refined target gene list with an integrative analysis that includes the microarray gene expression data. We apply our approaches to a large-scale ChIP-seq datasets in mESC (mouse Embryonic Stem Cell), and reconstruct the core transcriptional regulatory networks in this unique cell type.

Author's Publications Related to Thesis

Journals

- **Xu H.**, Handoko L., Wei X.L., Ye C., Sheng J., Wei C.L., Lin F., Sung W.K. (2010) “A Signal-noise Model for Significance Analysis of ChIP-seq with Negative Control”, *Bioinformatics*. 26:1199-1204.
- **Xu H.**, Wei C.L., Lin F., Sung W.K. (2008) “Genome-wide Identification of Differential Histone Modification Sites from ChIP-seq Data”. *Bioinformatics*, 24:2344-2349.
- **Xu H.**, Sung W.K., and Lin F. (2007) “A Statistical Approach for Identifying Differentially Expressed Genes in Time-course CDNA Microarray Experiment without Replicate”. *Journal of Bioinformatics and Computational Biology*, 5:281-296.

Book Chapter

- **Xu H.** and Lin F. (in press) “Computational analysis of ChIP-seq data and its application to embryonic stem cells”, in Zhan M. et al. (ed.) *Computational Biology of Embryonic Stem Cells*, Bentham Scientific Publishers.

Glossary

Absolute Error	the magnitude of the difference between the exact value and the approximation
aCGH	array comparative genomic hybridization
CCAT	a control based ChIP-seq analysis tool. Please also refer to the term “ChIP-seq”
ChIP	chromatin immuno-precipitation
ChIP-chip	chromatin immuno-precipitation followed by microarray
ChIPDiff	A software tool for identifying differential histone modification sites from ChIP-seq data. Please also refer to the term “ChIP-seq”
ChIP-PET	chromatin immune-precipitation with paired-end-tags
ChIP-SAGE	chromatin immuno-precipitation followed by serial analysis of gene expression.
ChIP-seq	chromatin immuno-precipitation followed by high-throughput sequencing
CIP	ChIP intensity profile. Please also refer to the term “ChIP”
CNV	copy number variation
DFMS	differential histone modification site
DNA	deoxyribonucleic acid
eFDR	empirical false discovery rate
EM	expectation maximization
Ensembl	a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute (in this thesis, this term refers to the gene database developed in the project)
ESC	embryonic stem cell

FDR	false discovery rate
Fold-Change	the ratio of two comparable measurements in biological experiments.
GFP	green fluorescent protein
GLTR	global identifier of target regions. A software tool for ChIP-seq analysis. Please also refer to the term “ChIP-seq”
GO	gene ontology
HCNS	highly conserved non-coding sequence
HMM	hidden Markov model
HMS	histone modification site
HTS	high throughput sequencing
LINE	long interspersed nuclear element
MACs	model based analysis for ChIP-seq. A software tool for ChIP-seq analysis. Please also refer to the term “ChIP-seq”
MEF	embryonic fibroblast
mESC	mouse embryonic stem cell
MGC	database of mammalian gene collection
Motif	a sequence pattern that is widespread and has, or is conjectured to have, a biological significance. In transcription regulation, the definition of a motif is specified as a short nucleotide sequence pattern that is frequently observed at the TFBS. Please also refer to the term “TFBS”
MTL	multiple TF-binding loci. Please also refer to the term “TF”
NPC	neural precursor cell
PCR	polymerase chain reaction
PET	paired-end-tags
pFDR	positive false discovery rate

PWM	position weight matrix
qPCR	quantitative reverse transcriptase polymerase chain reaction
QuEST	a kernel density estimator-based tool for ChIP-seq analysis. Please also refer to the term “ChIP-seq”
RefSeq	database of NCBI reference sequence
Relative Error	the ratio of the absolute error to the magnitude of the exact value. Please also refer to the term “absolute error”
RT-PCR	reverse transcriptase polymerase chain reaction
SICER	a spatial clustering approach for the identification of ChIP-enriched regions from ChIP-seq data. Please also refer to the term “ChIP-seq”
SISSRs	site identification from short sequence reads. A software tool for ChIP-seq analysis. Please also refer to the term “ChIP-seq”
SNP	single-nucleotide polymorphism
TF	transcription factor
TFBS	transcription factor binding site
Transcriptional Regulatory Network	a graphic representation of the relationship between transcription factors and their target genes.
TSS	transcription start site
WCE	whole cell extract

Chapter 1

Introduction

On benefit of the rapid growth of ultra-high-throughput sequencing technologies in recent years, ChIP-seq has become the main stream for the genome-wide study of protein-DNA interactions and histone modifications. Large amount of ChIP-seq datasets have been generated and published. The analysis of ChIP-seq data posed new challenges to the bioinformatics community. This thesis addresses the computational approaches for the identification of protein-DNA binding sites and histone modification sites from one or two ChIP-seq libraries. Based on the prediction of ChIP-enriched genomic sites, we reconstructed the key transcriptional regulatory network in mouse embryonic stem cells (mESC) from a large-scale ChIP-seq dataset.

1.1. Development of the ChIP-seq Technology

The protein-DNA interactions and histone modifications are crucial genomic events that regulate the gene expression in mammalian cells. Such event can be detected by the Chromatin Immunoprecipitation (ChIP) technique [1]. In ChIP experiments, an immune reagent (usually called antibody) specific for a DNA binding factor or a modified histone protein is used to enrich target DNA sites, to which the protein was bound in the living cell. The enriched DNA sites are then identified and quantified.

For the gigabase-size genomes of vertebrates, however, it has been difficult to make ChIP measurements that combine high accuracy, whole-genome completeness, and high binding-site resolution. Motivated by the goal of a complete understanding of the protein-DNA interactome and the epigenome, the ChIP technique has been integrated with high-throughput microarrays. This technology, named ChIP-on-chip or ChIP-chip, is able to detect hundreds to thousands of ChIP-enriched sites in a single experiment. In 2000, ChIP-chip was first successfully employed in the study on budding yeast *S. cerevisiae* [2][3][4], which has a smaller genome of 12.5M bps in relative to mammalian species. A few years later, it was applied to mammalian cells when whole-genome microarrays were available for gigabase-size genomes [5]. The major drawback of ChIP-chip lies on the fact that millions of probes need to be designed to cover the genome, which makes this technology expensive. Moreover, the sensitivity and resolution of signal detection on microarray data is limited since their noise level is relatively high.

Due to the drawbacks of ChIP-chip, some researchers sought alternative solutions by integrating ChIP technique with sequencing technologies. The idea is to determine the genomic location for a number of randomly sampled ChIP fragments, by sequencing one end or both ends of the fragments and aligning the sequence reads to the genome. When the sample size is large enough, the target sites are enriched by the aligned reads and hence are detectable. Two technologies, ChIP-SAGE [6] and ChIP-PET [7], were developed based on this idea. In their protocols, ChIP-SAGE sequences the 5' end of a DNA fragment, while ChIP-PET sequences both ends of a fragment and forms paired-end-tags (PETs). In comparison with ChIP-chip, the sequencing-based technologies are advantageous in two aspects: a) they do not rely on probe design, hence provides almost

unbiased whole-genome coverage; b) they reduced the experimental noise that could be introduced by the microarray imaging devices. In previous study, the author of this thesis and his colleagues employed ChIP-PET technique to reveal distinct genomic compartments defined by histone modifications H3K4me3 and H3K27me3 [30]. Yet, applications of ChIP-SAGE and ChIP-PET were limited at that time (years 2004-2006) due to the high-cost and low-capacity of sequencing technologies and their complex protocols.

In early 2007, Illumina Inc. launched the Solexa 1G genome analyzer, which is a milestone of the new generation ultra-high-throughput sequencing technologies. In subsequent years, the sequencing capacity was tremendously improved, which led to successful genome-wide applications in various areas such as genome re-sequencing and SNP association [8], mRNA expression [9], DNA methylation [10], and 3-dimensional DNA structure [11]. One of the major (also the earliest) applications of the ultra-high-throughput sequencing technologies is ChIP-seq, in which DNA sequencing is coupled with ChIP experiments to identify histone modification sites (HMSs) or transcription factor binding sites (TFBSs). Barski *et al.* first employed ChIP-seq approach to profile 20 histone lysine and arginine methylations in human CD4⁺ T cell [12]. Johnson *et al.* further adapted the ChIP-seq protocol to determine the TFBSs of NRSF protein in Jurkat T cell [13]. Research work by these two groups demonstrated the impressive power of ChIP-seq for the study of transcription factors and epigenetic markers, in consideration of sensitivity, specificity and resolution.

Conceptually, ChIP-seq is derived from ChIP-SAGE and ChIP-PET, and is of certain similarity to these technologies. In comparison with ChIP-SAGE and ChIP-PET, the

protocol of ChIP-seq has been significantly simplified. The flow chart of a ChIP-seq experiment is shown in Figure 1.1.

Figure 1.1: The flow chart of ChIP-seq (Image on courtesy of Illumina Inc.)

There are five steps in a ChIP-seq experiment:

- i) The chromatin in the nucleus is cross-linked and fractionated by sonication;
- ii) The DNA fragments bound by the target protein are enriched by ChIP process using specific antibody;
- iii) The ChIP-DNA fragments are subject to a one-step adaptor ligation, limited PCR amplification, and size selection;
- iv) The 5' ends of the DNA fragments are sequenced simultaneously using genome analyzer and image analysis pipeline;
- v) The sequence reads are aligned to the reference genome assembly so that the original genomic loci of the ChIP fragments are determined.

The size of the reads varies from 25 bps to 36 bps. A large majority of these short reads contain sufficient sequence information such that they are uniquely aligned to the genome. To date, one sequencing run of this procedure can generate up to one billion sequence reads. That is to say, the 3G-bps human genome can be covered by the reads generated in a single run, with more than 5 coverage fold.

With the huge amount of aligned reads, one may generate a comprehensive protein-DNA interaction profile that facilitates the identification of TFBSs and HMSs with high sensitivity and specificity. For example, Chen *et al.* and Marson *et al.* used ChIP-seq to identify more than 10,000 Nanog binding sites in mouse embryonic stem cells [14][15], 10 times more than those identified in the previous study [16]. Moreover, the high resolution of ChIP-seq profile is impressive. Johnson *et al.* claimed that the maximum distance between ChIP-seq peaks and protein-binding motif is approximately 50 bps in

their study [13], significantly shorter than the estimation from ChIP-chip (>100 bps). With such a high resolution, protein-binding motifs can be discovered effectively by computational methods.

In short, ChIP-seq has been recognized as a more powerful technology than the previous ones for the study of transcription factors and histone modifications.

1.2. ChIP-seq Analysis Tools in Literature

In a ChIP-seq experiment, millions of reads sequenced from a ChIP library are aligned to the genome. The histogram of read counts represents a genome-wide profile in which counts of read are overrepresented at the loci of interest (i.e. TFBSs or HMSs). The patterns of ChIP-seq profile may vary among different applications. For most ChIP-seq datasets on transcription factors, the profiles are typically with sharp peaks centered by the actual binding sites; while for some datasets on histone modifications, much more flat patterns are observed in profiling (Figure 1.2), suggesting that the histone modification enzymes may affect a number of consecutive nucleosomes. Accordingly, various application-dependent tools have been developed for identifying ChIP-enriched sites.

The first software tool for ChIP-seq analysis on TFBS identification was introduced by Johnson *et al.* [13]. They employed a straight-forward algorithm to locate the summit of the peaks in the profile. The variation of read counts due to the random sampling of

Figure 1.2: Different patterns of ChIP-seq profiling for CTCF, H3K4me3, H3K27me3, and H3K36me3 in proximal region of MYC gene in human K562 cell-line. Image captured from UCSC genome browser (<http://genome.ucsc.edu/>).

fragments could be further smoothed using the Gaussian density kernel, which was implemented in the tool of FSeq [17]. Since a ChIP-seq read is sequenced from one end of a ChIP-enriched DNA fragment, the alignment to the genome assembly is directional and can be in either sense or anti-sense direction. Indeed, the aligned read is oriented towards the actual binding site. In the light of this, several subsequent tools utilized the orientation information of reads to pursue a higher resolution of binding site prediction. FindPeaks [18] extended the reads towards their orientations to form “virtual fragments”. Overlapping “virtual fragments” were accumulated into the ChIP-seq profile, which better represents the binding peaks with improved resolution. QuEST [19] generated density profile for sense and anti-sense reads individually, and determined the binding

sites as the center between sense and anti-sense peaks. SISSRs [20] predicted the binding sites as transition position from sense reads to anti-sense reads. As reported in their corresponding publications, these approaches improved the resolution of peak prediction (measured as the median distance between predicted peaks to binding motif) from ~50 bps to ~20 bps.

Unlike those for transcription factor study, the ChIP-seq applications to histone modifications are mostly designed to reveal the ChIP-enriched genomic regions that may span from hundreds of bps to ~100k bps. Therefore, the predicted histone modification sites need to be defined in term of regions rather than individual peaks. Mikkelsen *et al.* proposed a sliding-window based approach to predict histone modification sites [21]. In each step of sliding, the significance of read counts in the window was tested against a random Poisson model. They also developed an alternative approach based on hidden Markov model (HMM), which is more suitable to handle larger continuous regions. Zang et al. defined “islands” as small ChIP-enriched bins and clustered the proximal islands into larger regions [22]. This idea was implemented in SICER, which seems to be less sensitive to the bin size in comparison to Mikkelsen’s windowing approach.

A common question in ChIP-seq analyses is how to control the specificity of prediction, which is usually measured by the false discovery rate (FDR). In the early ChIP-seq applications without a negative control, a random model of uniform distribution of DNA fragments was assumed for significance analysis. In that case, the expected number of random peaks can be estimated from either known distributions (e.g. Poisson distribution [18], Negative binomial distribution [23]) or randomly simulated dataset [14]. The FDR was then computed to be the ratio of the expected number of random peaks to that of

observed peaks above certain threshold. However, recent studies showed that the uniform random model is too ideal due to the impact of sequencing and mapping biases, chromatin structure and genome copy number variations [24][25]. Therefore, it has been suggested that a negative control library, which can be generated using non-specific antibody or input DNA, be used in a ChIP-seq experiment to reflect these intrinsic biases.

The most straight-forward approach to utilize the control library is to set an empirical threshold of fold-change against control, in order to filter out the falsely predicted peaks introduced due to intrinsic biases [14]. However, fold-change measurement could be too sensitive to the random sampling variation when the sequencing is not deep enough. Therefore, several other tools performed statistical tests to determine the significant peaks. In MACs, a random model of Poisson distribution was assumed, where the parameter of Poisson distribution is a variable and was estimated from the read counts in local region from the control library [24]. CisGenome [23] and PeakSeq [26] tested the significance of read enrichment against control using binomial distribution. Additionally, a novel approach was implemented by Tuteja et al. in their tool GLITR, in which the absolute read counts and the fold-change of a peak were mapped to a 2D space for FDR estimation [27]. These tools efficiently reduced the false positives caused by intrinsic biases and provided better estimation of FDR.

The above-mentioned ChIP-seq analysis tools are summarized in Table 1.1.

Table 1.1: A summary of ChIP-seq analysis tools for predicting ChIP-enriched sites.
(“yes/no” in the column of “with control library” indicates the tool can work with or without a control library)

Name or description of tool	site representation	with control library	Significance measurement	Reference
Johnson <i>et al.</i> 's peak detection	peaks	no	read counts	[13]
Fseq	peaks	no	kernel density estimates	[17]
FindPeaks	peaks	no	Poisson p-value	[18]
SISSR	peaks	no	read counts	[20]
Johnson <i>et al.</i> 's peak detection	peaks	no	read counts	[13]
Fseq	peaks	no	kernel density estimates	[17]
FindPeaks	peaks	no	Poisson p-value	[18]
MACs	peaks	yes	Poisson p-value	[24]
QuEST	peaks	yes/no	kernel density estimates	[19]
Mikkelsen <i>et al.</i> 's sliding-window approach	regions	no	read counts	[21]
Mikkelsen <i>et al.</i> 's HMM approach	regions	no	probability	[21]
CisGenome (one-sample)	regions	no	negative binomial p-value	[23]
CisGenome (two-samples)	regions	yes	binomial p-value	[23]
PeakSeq	regions	yes	binomial p-value	[26]
GLITR	regions	yes	number of nearest neighbor in 2D space	[27]
SICER	regions	yes/no	island score	[22]

1.3. Reconstruction of Transcriptional Regulatory Networks

Study on transcriptional regulatory networks has attracted special interest of the computational biology community for its crucial role in understanding cellular responses to internal and external signals through gene transcription and translation processes. A transcriptional regulatory network consists of a set of nuclear proteins called transcription factors (TFs), and their target genes whose expression states are directly altered in response to the DNA-binding events of the TFs. Enumeration of these network components is an essential step toward the creation of a framework for systematic analysis of the biological processes.

An algorithm for systematically and efficiently reconstructing the transcriptional regulatory networks should be able to discover new transcription factor – target gene regulatory connections and the logic of how the involving TFs regulate the target genes. Various methodologies have been proposed for constructing a gene network based on gene expression data from DNA microarrays. Akutsu *et al.* introduced Boolean networks for predicting the components in a small-scale network [67]; Hoon *et al.* employed a differential equation models to infer the network from time-course gene expression data [18]; and a number of publications have proposed the approaches to use Bayesian networks to determine the regulatory relationship between genes in different applications [69][70][71].

There are several drawbacks for the gene network estimation from microarray data. Firstly, while the gene network contains a large number of genes, the information contained in gene expression data is limited by the number of microarray probes, their

quality, the experimental design, noise level, and measurement errors. Secondly, it is difficult to distinguish the direct target genes of a TF from indirect targets unless special treatment is applied in the experiment. Recently, Hu *et. al.* reconstructed the transcriptional regulatory network of *S. cerevisiae* by inspecting the gene expression changes in the knock-out experiments on 263 TFs [72]. However, this method seems to be very costly for mammalian species, considering thousands of TFs in mammalian cells and a much more complex transcription mechanism that involves the combinatorial effect of multiple TFs. Thirdly, the microarray expression data do not provide the detailed information of TF binding, such as the location of binding site, the motif modules, and the chromatin structures influenced by binding events, which are crucial in elucidating the transcription mechanism at molecular level.

Due to the above limitations, people came to the methodology of combining microarray data and protein-protein interactions, protein-DNA interactions, binding site information, existing literature and so on, for the reconstruction of transcriptional regulatory networks [73][74]. Among them, genome-wide mapping of protein-DNA interactions seems very promising as they provide information to associate the TF binding events and the transcription of target genes, which vividly depict the regulatory procedures. Therefore, ChIP-seq, as the major technology to map the protein-DNA interactions in genome-wide scale, introduced a new insight for the reconstruction of transcriptional regulatory networks.

1.4. Objectives of the Thesis

As have been reviewed in the previous sections, a large amount of ChIP-seq datasets have been generated and published to date, which posed new challenges and opportunities to the bioinformatics researchers. The ChIP-seq analysis involves not only the prediction of ChIP-enriched sites, but also subsequent analysis procedures that lead to novel biological discoveries. Three major problems in ChIP-seq analysis are listed below:

- a) How to reliably identify the ChIP-enriched sites from ChIP-seq datasets that may be generated with variant experimental design?
- b) How to compare two or more ChIP-seq libraries generated from different cell-types, or with different cell treatments, in order to determine the genomic events and sequence elements that regulate differential gene expressions?
- c) How to extract biologically interesting information, such as transcriptional regulatory networks, from large-scale ChIP-seq datasets?

The goal of this thesis work is to provide solutions to these problems by developing novel computational approaches and methodologies.

1.5. Major Contributions

In respect to the problems given in section 1.4, three major contributions are to be presented in this thesis:

- a) We developed two software tools, Peak-finder [14] and CCAT [28], to identify the ChIP-enriched genomic sites. Peak-finder was designed to process a single

ChIP library, and CCAT was designed for the experiments with a ChIP library and a control one.

- b) We proposed a HMM-based approach to compare two histone modification libraries generated from different cell-types or with different treatments. This approach was implemented in the software ChIPDiff [29].
- c) By analyzing a large-scale ChIP-seq dataset that involves multiple transcription factors, we made significant progress in reconstruction of the transcriptional regulatory networks in mouse embryonic stem cells (mESC) [14]. Our discoveries include the protein-DNA binding motifs, transcription factor co-localizations, target regulated genes, and finally the core transcriptional regulatory networks.

1.6. Organization of the Thesis

This report consists of 6 chapters. The next chapter, Chapter 2, is focused on the prediction of protein-DNA interaction sites from a single ChIP library. We devised the algorithms based on our analysis methods for ChIP-PET, a precursor of the ChIP-seq technology. The algorithms are integrated into the software Peak-finder [14].

In Chapter 3, we propose a signal-noise model of ChIP-seq experiment [28]. Based on this, we developed a general-purpose tool, CCAT (Control based ChIP-seq Analysis Tool), by which a control library is utilized to improve the performance of site identification. We show that CCAT significantly outperformed previous methods in

terms of sensitivity, and it predicted a number of weak histone modification sites that are of biological interest.

We introduce a HMM based approach to predict differential histone modification sites (DHMSs) in Chapter 4. This approach was implemented in the software ChIPDiff [29], which was designed to compare two histone modification libraries generated from different cell-types or with different cell treatments.

In Chapter 5, we employ a large-scale ChIP-seq dataset that involves 15 transcription factors or regulators in mESC [14], to demonstrate our analysis approaches that lead to the reconstruction of transcriptional regulatory network.

Finally, in Chapter 6, we summarize the achievements in the thesis work, discuss the limitations of the methods, and propose a few potential directions of future work.

Chapter 2

Single Library Analysis

This chapter describes the algorithms for identifying ChIP-enriched sites from a single ChIP-seq library, which was implemented in the tool Peak-finder [14]. The algorithms we developed were derived from the previous ChIP-PET analysis methods [30]. Therefore, we first introduce a peak-finding algorithm developed for ChIP-PET analysis; then, we show how the ChIP-PET analysis methods are adapted to ChIP-seq datasets based on the concept of “virtual fragment”; next, the peak-calling approach is discussed in detail; and finally, we employed real ChIP-PET and ChIP-seq datasets to demonstrate the performance of the proposed approaches.

2.1. Peak-finding for ChIP-PET Analysis

ChIP-PET (Chromatin Immuno-Precipitation Paired End Tags) was first introduced in 2006 to identify the genome-wide p53 binding sites in a cancer cell-line [7]. The procedure of a ChIP-PET experiment is shown in Figure 2.1a. In ChIP-PET protocol, both ends of a ChIP DNA fragment are sequenced by ligating the ends and generating tags in pairs. The PET sequences are mapped to genome to demarcate the boundaries of DNA fragments. PET singletons are considered background, while overlapping PETs (PET clusters) are regarded as enrichment by the same immune-precipitation events. The overlapping regions (PET overlap) in a PET cluster may, therefore, contain the ChIP-enriched sites. In the earlier processing pipeline proposed by Wei *et al.* [7], a PET cluster

(a)

(b)

(c)

Figure 2.1: (a) The procedure of a ChIP-PET experiment (Image on courtesy of Cell Press). (b) Generating ChIP Intensity Profile (CIP) from PETs. (c) Estimation of the distribution of fragment size. Curve in blue represents the distribution of fragment length, and curve in red represents the fitted gamma distribution density function. The p-value of Kolmogorov-Smirnov goodness-of-fit test is 0.187.

is assigned a confidence score which is the number of PETs in the cluster. A PET cluster with k PETs is called a PET k . Monte Carlo simulation is employed to determine the cutoff threshold. The limitation of the PET clustering method is that a PET cluster spans thousands of base pairs, for which the resolution is comparatively low.

To achieve a higher resolution, we proposed a peak-finding algorithm to locate the binding sites from the CHIP Intensity Profile (CIP). A CIP is generated by assigning each genomic locus with an intensity score which equals to the number of PETs overlapping this locus, where a binding site is expected to be located at a peak of a CIP (Figure 2.1b). A peak is defined as the locus with the maximum intensity in a window centered by that locus. If there are more than one maximum in the window, the one to the 5' is selected as the peak. The peak finding algorithm is as following:

- i) Select all the local maximums, without window restriction, in the CIP. This generates a list of candidate peaks.
- ii) For each candidate peak in the list, search the region by half window size to the 5'. If there is any locus with higher or equal intensity in the region, remove the candidate peak from the peak list.
- iii) For each candidate peak remained in the list, search the region by half window size to the 3'. If there is any locus with higher intensity in the regions, remove the candidate peak from the peak list.
- iv) All the candidate peaks are ranked in descending order of their intensity. Peaks with intensity larger than a threshold are reported. In [30], we determine the threshold based on quantitative RT-PCR (qPCR) validation.

In the peak finding algorithm, if the window size is too large, a true binding site adjacent to another higher intensity bind site will be missed; while if the window size is too small, a trivial peak caused by noisy PETs around a true binding site will be deemed as binding site. To determine the window size, we first estimate the PET length distribution from all PETs. Next, the PET length distribution is fitted with a gamma distribution density function $g(x; \alpha, \beta)$ [31][32], where x is a random variable from gamma distribution, α is the shape parameter, and β is the scale parameter:

$$g(x; \alpha, \beta) = x^{\alpha-1} \frac{\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)} \quad (2.1)$$

The assumption of gamma distribution is based on Qi *et al.*'s experiment, in which the lengths of randomly sheared DNA fragments can be fitted with a gamma function (Figure 2.1c) [33]. Based on the mathematical characteristic of the gamma distribution, the distribution of the distance between a binding site to an arbitrary end of a fragment captured from the binding site can be easily represented by: $g(x; \alpha/2, \beta)$. We determine the window size as two times the 90 percentile in $g(x; \alpha/2, \beta)$, so that 90% of the PETs overlapping a peak will not contribute to the intensity of the peaks adjacent to that peak. By this way, we guarantee that majority of the peaks in our list are non-trivial.

2.2. ChIP-seq Profiling with Virtual Fragment

In a ChIP-seq experiment, only the 5' end of a ChIP fragment is sequenced, and the genomic location of the 3' end is unknown. Therefore, the reads mapped to the positive strand of the DNA assembly (i.e., sense alignment) are concentrated in the upstream of a binding site; while the reads mapped to the negative strand (i.e., anti-sense alignment) are concentrated in the downstream (Figure 2.2a). To adapt the peak-finding algorithm to ChIP-seq applications, we introduced a concept of “virtual fragment”. The idea is to approximate the genomic locations of 3' ends based on the fragment length distribution, hence the virtual fragments can be generated by linking the aligned 5' ends with the approximated 3' ends.

We first estimate the fragment length distribution. Suppose a library contains m positive orientation tags $\{t_1^+, t_2^+, \dots, t_m^+\}$ and n negative orientation tags $\{t_1^-, t_2^-, \dots, t_n^-\}$. For any pair of tags (t_i^+, t_j^-) , there are two alternative situations: (a) the corresponding DNA fragments are captured from the same binding site for both tags; (b) the DNA fragments are captured from different binding sites, or at least one fragment is unbound by the target protein. For the situation (a), the distance $d(t_i^+, t_j^-)$ follows the same fragment length distribution, which is assumed to be a gamma distribution; for the situation (b), the two tags are independent and distribution of the distance $d(t_i^+, t_j^-)$ is assumed to be uniform. Therefore, the distribution of $d(t_i^+, t_j^-)$ is expected to be a mixture of a gamma distribution and a uniform distribution. The algorithm for estimating the fragment length distribution is detailed as following:

- i) For each t_i^+ , $i = 1, 2, \dots, m$, search for all the negative orientation tags within l bps, denoted $\{t_j^-, t_{j+1}^-, \dots, t_{j+k}^-\}$, where l is the maximum length of a fragment, and k is the number of the negative orientation tags satisfying the criterion;
- ii) Calculate the pair-wise distances $\{d(t_i^+, t_j^-), d(t_i^+, t_{j+1}^-), \dots, d(t_i^+, t_{j+k}^-)\}$ for all t_i^+ , $i = 1, 2, \dots, m$, and plot the distribution of the distances;
- iii) The distribution calculated in ii) is fitted by the function:

$$f(x; \lambda, \alpha, \beta) = (1 - \lambda)\pi_l + \lambda g(x; \alpha, \beta) \quad (2.2)$$

where π_l is the density function of a uniform distribution with span equal to l , which represents the distribution of distance of two randomly paired tags; $g(x; \alpha, \beta)$ is a gamma function defined in Eq. 2.1 and depicts the fragment length distribution; λ is the weight coefficient in the linear combination.

A simple approach to approximate the 3' of a fragment is to extend the 5' tag by d base-pairs towards its orientation, where d was determined based on the fragment length distribution $g(x; \alpha, \beta)$. In our implementation, d was set to be the 90 percentile of the gamma distribution $g(x; \alpha, \beta)$, such that 90% of the virtual fragments would contribute to the ChIP intensity at the actual binding site. The virtual fragments were then accumulated to form the CIP profile as described in Section 2.1. Figure 2.2b gives an example of the virtual fragments and the CIP generated from the virtual fragments.

A drawback of the above method is that 10% of the virtual fragments do not contribute to the actual binding sites. Therefore, we developed an improved approach to generate the

CIP from ChIP-seq data in the later versions of the software. The idea is to assign the tags to a genomic location based on a weight derived from the gamma distribution $g(x; \alpha, \beta)$. Suppose there are i sense tags denoted $\{t_1^+, t_2^+, \dots, t_i^+\}$, and j anti-sense tags denoted $\{t_1^-, t_2^-, \dots, t_j^-\}$, within the maximum distance l from a genomic location x , the ChIP intensity at x is represented as:

$$CIP(x) = \sum_{k=1}^i g(x - t_k^+; \alpha, \beta) + \sum_{k=1}^j g(t_k^- - x; \alpha, \beta) \quad (2.3)$$

By this, all the tags contribute the ChIP intensity in their proximal regions hence the utilization of the data is improved.

This improved method, namely “weighted virtual fragment”, was compared to the previous method of fixed-length virtual fragment, using the NRSF data published in [13] and the CTCF data published in [14]. In the comparison, the peak-finding algorithm proposed in Section 2.1 was applied to the CIP generated by each method; the top 1000 peaks with the highest intensity were selected; and the average distance of the peak locus to the nearest binding motif was computed. The motifs were employed in this comparison since they define the exact protein-binding genomic positions, therefore the distance between the motif and the predicted peak reflects the precision of binding site prediction. As shown in Figure 2.3, the weighted virtual fragment method significantly improved the precision in both NRSF and CTCF experiments.

(a)

(b)

Figure 2.2: (a) Sense and anti-sense tags are distributed in different sides of the binding site; (b) a demonstration of virtual fragments and the ChIP Intensity Profile (CIP) of ChIP-seq.

Figure 2.3: The weighted virtual fragment method improves the precision of binding site prediction.

2.3. Peak-calling

The peak-finding algorithm proposed in Section 2.1 reports a list of candidate peaks. These peaks are ranked in descending order of their ChIP intensities. Which peaks should be selected to represent the reliably predicted binding sites? We employed a peak-calling strategy for this purpose.

We first introduce a random model of fragment distribution. In the random model, we assume the fragments are captured without ChIP-enrichment by the antibody. Therefore,

the fragments are uniformly distributed in the genome by random expectation. For ChIP-PET analysis, we performed Monte-Carlo simulation [34] to uniformly place the PETs in the genome, and performed the peak-finding algorithm on the simulated data to determine a list of random peaks. The observed peaks are then compared to the random peaks based on the distribution of peak intensity. Given a threshold of intensity, the expected false positive rate can be computed as the ratio of the number of random peaks to the number of called peaks. Table 2.1 compared the observations against the simulation results for p53 ChIP-PET data published in [7]. As shown in the table, when the peak intensity is larger than or equals to 3, we expect a false positive rate <1%, therefore a threshold of 3 can be determined for peak selection. Similarly, ChIP-seq peaks are called based on simulation data in which the tags, rather than the PETs, are uniformly placed in the whole genome.

Monte-Carlo simulation is a reasonable approach for peak-calling when only a single ChIP-seq library is available. However, simulation could not reflect the intrinsic biases of

Table 2.1: Comparison of p53 peaks identified in ChIP-PET experiment [7] against simulation results.

	total PETs	ChIP intensity					
		1	2	3	4	>=5	>=3
#peaks in experiment	65,572	61,270	2,886	471	252	693	1,416
#peaks by simulation	65,572	64,790	770.9	11.1	0.0034	<0.0001	11
expected false positive rate			0.267117	0.023567	1.35E-05	0	0.007768

ChIP-seq, such as the sequence mappability, chromatin structure, and DNA copy number variations [24][25]. When a control library is available, the predicted peak list can be further refined by comparing the ChIP library against the control. Although a more complex approach for ChIP-control comparison will be proposed in Chapter 3, herein we first introduce a naïve method based on fold change computation. This strategy was employed in our Peak-finding software package to enhance its generality. Given a predicted peak x with the intensities $I_{x,obs}$ in ChIP profile and $I_{x,control}$ in control profile, the fold-change at x is represented as:

$$FC(x) = \frac{I_{x,obs} + \varepsilon}{I_{x,control} + \varepsilon} \times \frac{N}{M} \quad (2.4)$$

where M and N are the total number of tags (or PETs for ChIP-PET) in the ChIP and control libraries, respectively. The parameter ε is a positive constant called pseudo-count and is usually set to be 1. By this definition, the fold-change is normalized against the sequencing depth of the libraries, and the peaks with a fold-change smaller than a threshold are filtered out.

2.4. Application to ChIP-PET Histone Modification Dataset

We applied the ChIP-PET peak-finding algorithm to the human embryonic stem cell (ES cell) H3K4me3 and H3K27me3 histone modification ChIP-PET data [30]. Histone modifications are crucial for proper lineage specification and embryo development. Among them, H3K4 and H3K27 trimethylation are two histone modifications associated with transcriptional active and repressed regions, respectively. To determine the impacts of these marks on chromatin organization and in regulating the gene expression programs

in human ES cells, we profiled the whole genome H3K4me3 and H3K27me3 modifications by ChIP coupled with ChIP-PET multiplex sequencing strategy.

To achieve comprehensive coverage, a total of 2.1 million and 3 million PET sequences were generated from H3K4me3 and H3K27me3 libraries, respectively. These PETs defined 679,765 and 992,534 unique ChIP fragment mapping locations, respectively. These PET locations are processed by the pipeline. With an expected false positive rate <1%, we identified 17,167 H3K4me3 modification peaks and 10,554 H3K27me3 peaks. RT-PCR wet-lab validation showed that ~100% of these peaks are real histone modification sites.

Because H3K4me3 is known to associate with the 5' regions of actively transcribed genes [35][36][37], we first examined the correlation between these 17,168 PET-defined H3K4me3 loci and transcription start sites (TSS) of known genes. Consistent with significant correspondence between H3K4me3 sites and TSSs in other cell types, vast majority (14,325; 83.4%) of our 17,168 H3K4me3 were located within 1 kb of TSS of well annotated Refseq genes.

We next examined the genes associated with each type of histone modification. Taking from the 19,247 well annotated human Refseq genes located on chromosome 1-22, 12,829 (67%) Refseq genes contain H3K4me3 and only 1,956 (10%) Refseq genes contain H3K27me3 within ± 1 Kb of their TSS (Figure 3A). Surprisingly, among these 1,956 H3K27me3 modified genes, almost all (1,899; 97%) of them contained H3K4me3 peaks within ± 1 Kb of their TSS and vast majority (1,746; 90%) of them harbored H3K4me3 peaks.

Based on the statistics above, the entire genes in ES genome can be categorized into three distinct compartments; 11,083 of them contain only H3K4me3, 1956 of them contain both H3K4 and K27me3 and 6208 of them contain neither modification (Figure 2.4). Furthermore, tight associations exist between the genes in each compartment and their transcriptional activities. To explore such distinct compartments in relation to the unique identity of human embryonic stem cells, namely pluripotency and self-renewal, we characterized the functions of these genes based on Gene Ontology (GO) categorization through PANTHER database (<http://www.pantherdb.org>). Specific molecular activities and biological processes that were exclusively enriched in different sections were revealed. Details of gene ontology associated with H3K4me3 and H3K27me3 marks are given in Appendix I.

Figure 2.4: The Venn diagram of H3K4me3/H3K27me3 marked genes reveals distinct genomic compartments in human ES.

2.5. Application to ChIP-seq Transcription Factor Dataset

In this section, we employ Chen *et al.*'s dataset [14] in mESC to demonstrate the performance of Peak-finding in identifying the transcription factor binding sites (TFBSs) from ChIP-seq data. The ChIP-seq dataset consists of 16 libraries, including 13 sequence-specific TFs (Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, and CTCF), 2 transcription regulators (p300 and Suz12), and a negative control (GFP). The 15 factors in the experiment design are known to play different roles in ES-cell biology as components of the LIF and BMP signaling pathways, self-renewal regulators, and key reprogramming factors. The libraries were prepared on E14 mouse ES cells.

We applied Peak-finding software to these 15 TF libraries. An FDR cut-off of 1% was set for most libraries except for Smad1 and P300, for which relatively less number of peaks were reported hence we used an FDR cut-off of 5% in order to include more peaks for subsequent analysis. The fold-change threshold was set to be 5-folds.

As the result, we identified hundreds to tens of thousands binding peaks for different libraries. The library information and prediction results are summarized in Table 2.2. To validate the reliability of predictions, we designed quantitative RT-PCR (qPCR) experiments for the peaks randomly selected from the predicted list. More than 30 peaks were selected for each library in the validation experiment. As shown in Table 2.2, the predictions are highly reliable with the positive rate larger than 90%, indicating that the approaches proposed in this Chapter performed reasonably well on the dataset. Interestingly, the enrichment fold-change measured by qPCR validation is almost linearly

correlated with the ChIP intensity of the peaks (Figure 2.5), implying that ChIP-seq analysis not only renders qualitative representation of binding sites, but also provides quantitative information of real binding affinity. Further analysis of this dataset will be described in Chapter 5.

Table 2.2: A summary of library information of the mouse ES transcription factor dataset [14], and the corresponding peak prediction results.

Factor name	Number of total uniquely aligned reads	Number of predicted peaks	FDR	Positive rate by qPCR validation (> 3 fold)
Nanog	8,424,102	10,343	0.01	100%
Oct4	4,911,144	3,761	0.01	92%
Sox2	4,821,446	4,526	0.01	100%
Smad1	3,338,896	1,126	0.05	96%
p300	5,424,357	524	0.05	90%
E2f1	8,449,181	20,699	0.01	100%
Tcfcp2l1	8,787,961	26,910	0.01	92%
Suz12	5,986,520	4,215	0.01	100%
CTCF	3,686,056	39,609	0.01	95%
Zfx	3,844,429	10,338	0.01	100%
Stat3	5,351,116	2,546	0.01	97%
Klf4	3,807,970	10,875	0.01	95%
n-myc	4,823,212	7,182	0.01	100%
c-myc	6,637,404	3,422	0.01	95%
Esrrb	7,148,528	4,7888	0.01	88%
GFP (control)	3,578,466	-	-	-

Figure 2.5: ChIP-seq intensity is almost linearly correlated with qPCR fold-change measurement. Data retrieved from the Nanog library in Chen *et al.*'s dataset [14]. The Pearson correlation is 0.947.

Chapter 3

Comparing ChIP Library with Control

In Chapter 2, we proposed a peak-finding based approach to identify the ChIP-enriched peaks from a single ChIP-seq library. However, the ChIP-seq reads are subject to intrinsic biases due to the sequence mappability, chromatin structure and genome copy number variations [24][25]. These biases result in significant false positives in the prediction from a single library. Therefore, a control library has been suggested to be included in ChIP-seq analysis [23][24][26]. In this chapter, we first discuss the source of intrinsic biases of ChIP-seq. Next, we propose a signal-noise model, in which a concept of “noise rate” was introduced to measure the noise level of ChIP-seq data. We proposed an iterative algorithm to estimate the noise rate, and derived a library-swapping strategy from the signal-noise model for the FDR estimation. Our approaches are integrated in a general-purpose framework, named CCAT (Control based ChIP-seq Analysis Tool), for the significance analysis of ChIP-seq. Finally, we applied CCAT to the H3K4me3 and H3K27me3 dataset on mouse embryonic stem cells (mESCs) [21]. Our result showed CCAT significantly improved the sensitivity of site prediction. Moreover, the weak ChIP-enriched sites predicted by CCAT are biologically meaningful.

3.1. Intrinsic Bias of ChIP-seq

In early ChIP-seq application without a negative control, reads are assumed to be uniformly distributed in the genome in a random background model [14][18]. However,

recent studies showed that the uniform model is too ideal due to the existence of intrinsic biases of ChIP-seq read enrichment. There are three major sources of intrinsic biases:

- **Sequence mappability.**

When aligning the ChIP-seq reads to the genome assembly, some reads are not mappable due to two reasons:

- (i) The genome assembly is not complete and the reads captured from unsequenced DNA are discarded.
- (ii) Segmental duplication is common in mammalian genomes. Reads captured from segmental duplication regions may be aligned to multiple genomic loci hence are excluded in generating ChIP intensity profile.

This situation of (ii) is often observed in the repetitive regions. As an example, Figure 3.1a shows that the ChIP-seq reads are depleted in a LINE (Long Interspersed Nuclear Elements) repeat region in multiple libraries published in [21].

- **Chromatin structure.**

Chromatin is the complex combination of DNA and proteins that forms chromosomes. It is divided between heterochromatin (condensed) and euchromatin (extended) forms [38]. The euchromatin is enriched by genes and active regulatory elements, while the heterochromatin is mainly in the gene desert. Recently, evidence from different groups indicated that the ChIP-seq reads are intrinsically enriched around the transcription start sites (TSSs) [25][26], suggesting bias associated with chromatin structure. To validate this argument, we employed a WCE (Whole Cell Extract) library prepared on mESC. In

the WCE experiment, no antibody was applied, therefore the distribution of reads is intrinsic to the ChIP-seq technology. We aligned the TSS of RefSeq genes, and plotted the cumulative frequency of WCE reads in relative to TSS. As shown in Figure 3.1, an obvious peak around the TSS was observed (Figure 3.1b). Our result, together with previous evidence, confirmed that the bias of ChIP-seq reads is associated with chromatin structures.

- **DNA copy number variation.**

Most mammalian species ordinarily have two copies of each autosomal region, one per chromosome. This may vary for particular genetic regions due to deletion or duplication. A copy number variation (CNV) is a segment of DNA in which copy-number differences have been found by comparison of two or more genomes [39]. When the copy number of a DNA segment is larger than two, more ChIP-seq reads are captured from the corresponding loci in relative to other genomic regions. This results in the enrichment of reads in the DNA segments with larger copy number. Vega *et al.* [25] compared the ChIP-seq profile of a WCE library with array CGH (aCGH) data in cancer cell-line MCF-7. The array CGH data was used as the benchmark for copy number variation. They reported significant correlation of the WCE profile and aCGH data (Figure 3.1c), which indicates that the DNA copy number variation is an important source of ChIP-seq bias.

(a)

(b)

Figure 3.1 (to be continued): Source of intrinsic biases of ChIP-seq (to be continued). (a) depletion of ChIP-seq reads in LINE repeats; (b) enrichment of reads in relative to TSS. Black and blue curves correspond to the intensity of sense and anti-sense reads, respectively.

(c)

Figure 3.1 (continued): Source of intrinsic biases of ChIP-seq. (continued) (c) the WCE ChIP-seq intensity is correlated with aCGH data in MCF-7 cell-line.

3.2. The Signal-noise Model of ChIP-seq

In a ChIP-seq experiment, majority of the unbound DNA fragments are washed out in the immune-precipitation procedure. The ChIP-processed library is enriched by the fragments pulled down from the genomic loci with high chance of protein-DNA interactions or histone modifications. However, considerable “non-useful” fragments remain in the library due to the random protein-DNA or antibody-DNA contacts that are not position-specific. Reads sequenced from these fragments are widely spread in the genome, and are considered background noise in addition to the real signal of ChIP enrichment.

In ChIP-seq analysis, the biologically interesting sites are identified by searching for the genomic loci where the reads sequenced from ChIP DNA are over-represented in relative to the background with no ChIP enrichment. In literature, a number of ChIP-seq analysis

tools have been developed for this purpose [18-27]. All these tools perform well to extract strong ChIP-enriched sites. However, two questions remain to be answered: (a) to which extent is a ChIP-seq experiment able to reveal the weak ChIP-enriched sites? (b) are the weak sites biologically meaningful? To answer these questions, it is necessary to establish a signal-noise model to separate the weak ChIP signals from background noise.

Before proposing the signal-noise model, we first model the random sampling process in a ChIP-seq experiment, in which millions of reads were randomly retrieved from the whole genome. Suppose N reads were retrieved from a genome of length L , a proper analogy is the case where N balls were randomly drawn from L boxes. Let the random variable s_i indicate the number of reads observed at the genomic location i . This is mathematically equivalent to observing s_i balls drawn from the i th box in the ball-drawing problem. Therefore, the vector $S = [s_1, s_2, \dots, s_L]^T$ follows a multinomial distribution with parameters P and N , where $P = [p_1, p_2, \dots, p_L]^T$ and $\sum_{i=1}^L p_i = 1$.

When a specific genomic locus i is concerned, the observation s_i follows the single-component marginalized multinomial distribution, which is binomial with parameters N and p_i . Since $p_i \ll 1$ (note genome size is large) and N is very large in the ChIP-seq application, the binomial distribution of s_i can be closely approximated using a Poisson distribution with a single parameter $\lambda = Np_i$, which is mathematically more concise for further analysis.

In short, the multinomial, binomial, and Poisson models proposed here are equivalent for the random sampling process of ChIP-seq [18][23].

Now let's discuss about the signal-noise model. Considering a control library with M uniquely aligned reads that were mapped to a genome of length L , the control profile can be represented as a vector of observations $[c_1^+, c_1^-, \dots, c_L^+, c_L^-]^T$, where c_x^θ refers to the number of control reads mapped to the genomic location x in direction θ ($x \in [1, L]$; $\theta \in \{+, -\}$). Due to the random sampling process in ChIP-seq experiment and the rationale given above, c_x^θ approximately follows a Poisson distribution:

$$c_x^\theta \sim \text{Poisson}(u_x^\theta M) \quad (3.1)$$

where u_x^θ is a normalized factor such that $\sum_{x,\theta} u_x^\theta = 1$.

Next we consider a ChIP library with N uniquely aligned reads. Similarly, the number of ChIP reads mapped to the genomic location x in direction θ , denoted d_x^θ , follows a Poisson distribution:

$$d_x^\theta \sim \text{Poisson}(v_x^\theta N) \quad (3.2)$$

where v_x^θ is a normalized factor such that $\sum_{x,\theta} v_x^\theta = 1$. We modeled v_x^θ to be the linear combination of the signal s_x^θ and the noise n_x^θ , such that:

$$v_x^\theta = (1 - \alpha)s_x^\theta + \alpha n_x^\theta \quad (3.3)$$

where $\sum_{x,\theta} s_x^\theta = 1$ and $\sum_{x,\theta} n_x^\theta = 1$. α is the weight coefficient in the linear combination.

To employ the control library as independent observations of background noise, we made an assumption that the normalized control vector $[u_1^+, u_1^-, \dots, u_L^+, u_L^-]^T$ and the noise vector $[n_1^+, n_1^-, \dots, n_L^+, n_L^-]^T$ follow the same multivariate prior distribution π , which models the

intrinsic read bias of ChIP-seq. The variation of π reflects slight differences of read bias between ChIP and control libraries, which could be introduced to the experiment prior to random sampling. Note that π is conditional on the chromatin structure, mapping bias, and copy number variations. Therefore, the ChIP and control libraries need to be generated on the same cell-type, with the same chromatin preparation procedure and identical configurations of alignment software. A line of evidence supporting the above assumption was given by Rozowsky *et al.* [26]. They generated scatter-plot of ChIP and control read counts on a list of genomic regions with no detectable ChIP enrichment. As the result, a linear and nearly symmetric scatter pattern was observed.

For the convenience of description, we denote E to represent the set of genomic locations with real ChIP signal, hence $\bar{E}$ refers to the background with $s_x^+ = 0$ and $s_x^- = 0$. Based on the signal-noise model and the assumption of intrinsic read bias, we have the following proposition:

Proposition 1: Given a genomic region $r = [a, b]$, and $r \subset \bar{E}$, the observation vectors $\mathbf{c}_r = [c_a^+, c_a^-, \dots, c_b^+, c_b^-]^T$ and $\mathbf{d}_r = [d_a^+, d_a^-, \dots, d_b^+, d_b^-]^T$ follow the same multivariate distribution under the hyperprior π , if $M = \alpha N$.

The proof of Proposition 1 is given in Appendix II of this thesis. Proposition 1 indicates that the read counts from the control library provide unbiased measurements of the noise level in the ChIP library when the sample sizes satisfy the condition $M = \alpha N$. In this context, the parameter α is called noise rate, which ranges from 0 to 1 and refers to the fraction of noise in the ChIP library. Theoretically, the noise rate is associated with the normalization factor r_0 in Ji *et al.*'s approach [23], such that $\alpha = r_0 M / N$. Generally

speaking, a smaller noise rate implies better ChIP quality in the experiment. In the worst situations where the noise rate approaches 1.0, all the ChIP reads are noise hence no ChIP-enriched site is detectable. Therefore, estimation of the noise rate provides an assessment of the data quality, and further facilitates the subsequent significance analysis.

3.3. Noise Rate Estimation

If a set of background regions R is known in prior and there is sufficient reads in R , the noise rate can be approximated as the ratio of ChIP read counts to control read counts in these regions, normalized against the sequencing depths:

$$\alpha \approx \frac{\sum_{r \in R} \sum_{x \in r, \theta} d_x^\theta}{\sum_{r \in R} \sum_{x \in r, \theta} c_x^\theta} \times \frac{M}{N} \quad (3.4)$$

Unfortunately, a predefined background region set is unavailable for most ChIP-seq applications. Ji *et al.* [23] selected the background to be the regions with small read counts. However, due to the intrinsic bias of ChIP-seq, the read counts may also be small for some ChIP-enriched regions and may be relatively large for some background regions. Therefore, a better solution is to determine the background by comparing the ChIP and control libraries, rather than by using the absolute read counts. In our approach, we first partitioned the whole genome into non-overlapping 1kb bins. The background regions were then selected to be the bins with ChIP read counts less than the expected noise read counts estimated from the control. That is:

$$R^* = \{r : \sum_{x \in r, \theta} d_x^\theta < \alpha N(\sum_{x \in r, \theta} c_x^\theta) / M\} \quad (3.5)$$

Based on Eq. 3.4 and 3.5, we proposed an iterative algorithm for estimating noise rate. In our algorithm, the ChIP and control data was divided into two subsets: D^+ consisting of sense reads, and D^- consisting of antisense reads. All the reads were shifted by $l/2$ bps towards their orientations, where l is the average DNA fragment length and could be approximated either experimentally or computationally [18][24]. In each step of iteration, a set of background regions were selected using D^+ , followed by updating noise rate using D^- . The iterative algorithm is described as follows:

Initialization:

Divide the datasets into D^+ and D^- ; Partition the whole genome into non-overlapping 1kb bins; Set $\alpha_0 = 1.0$;

The i th iteration:

- i) Count the reads from D^+ for the bins, and select a set of background bins R_i^* based on Eq. 3.5, where $\alpha_0 = \alpha_{i-1}$;
- ii) Count the reads in D^- for the bins in R_i^* and compute α_i based on Eq. 3.4;

Termination:

$\alpha_i > \alpha_{i-1}$, or i larger than a threshold.

In our test with real datasets, the above algorithm converged very fast within 10 iterations (Figure 3.2).

Figure 3.2: Convergency of the iterative noise rate estimation algorithm on H3K4me3 and H3K36me3 datasets [21].

3.4. FDR Estimation by Library Swapping

Several tools have been proposed in the literature for ChIP-seq analysis with negative control [19][23][24][26][41]. In general, these tools perform three steps: (i) select candidate sites; (ii) rank sites based on certain significance measurement; and (iii) determine the cutoff threshold. Most methods predict ChIP-enriched sites based on the read counts in a local region. For these cases, steps (i) and (ii) can be represented as a function f that maps the read count vectors of a local region to a significance measure. Preferentially, the threshold in step (iii) is determined with FDR control. In this thesis, we take the form of Storey's pFDR (also called Bayesian FDR) [42]. Storey proposed that the pFDR can be represented as a conditional probability:

$$pFDR(\Gamma) = \Pr(H = 0 \mid T \in \Gamma) \quad (3.6)$$

where H is the null hypothesis, T is a significance measure, and Γ is the rejection region. In our application, given a list of non-overlapping regions, a significance measure f , and a threshold t for rejection, the pFDR can be written as:

$$pFDR(t) = \Pr(r \subset \bar{E} \mid f(\mathbf{c}_r, \mathbf{d}_r) > t) \quad (3.7)$$

where r represents a genomic region, $\bar{E}$ is the union of background regions, and $\mathbf{c}_r$ and $\mathbf{d}_r$ are the corresponding read count vectors, as defined in Proposition 1.

Based on Proposition 1 and Eq. 3.7, we have:

Proposition 2: Under the conditions:

- a) $\Pr(f(\mathbf{d}_r, \mathbf{c}_r) > t \mid r \notin \bar{E}) \ll \Pr(f(\mathbf{c}_r, \mathbf{d}_r) > t \mid r \notin \bar{E})$, and
- b) $M = \alpha N$;

The pFDR for a threshold t can be approximated by:

$$pFDR(t) = \frac{\#\{f(\mathbf{d}_r, \mathbf{c}_r) > t\}}{\#\{f(\mathbf{c}_r, \mathbf{d}_r) > t\}} \quad (3.8)$$

where the operator $\#\{\Delta\}$ is defined to be the total number of regions satisfying the condition Δ .

The proof of, and special notes on Proposition 2 are given in Appendix III of this thesis. Proposition 2 underlies a library-swapping strategy of FDR estimation, in which the null distribution of significance measurement is empirically estimated by exchanging the ChIP and control libraries, i.e., control versus ChIP. The library-swapping strategy was first introduced by Zhang *et al.* [24] in the software MACs. However, as mentioned by

Zhang *et al.*, the estimated FDR would be biased if the read sample size from ChIP and control are not balanced. In Proposition 2, we give two necessary conditions for the correctness of library-swapping strategy.

The first condition indicates that the significance function f need to be defined such that the distributions of $f(\mathbf{c}_r, \mathbf{d}_r)$ and $f(\mathbf{d}_r, \mathbf{c}_r)$ are well-separated for the ChIP-enriched regions. For example, if we define $f(\mathbf{c}_r, \mathbf{d}_r) = \|\mathbf{c}_r\| + \|\mathbf{d}_r\|$ ($\|\bullet\|$ is the magnitude operator), the library-swapping strategy is not applicable since the distributions of $f(\mathbf{c}_r, \mathbf{d}_r)$ and $f(\mathbf{d}_r, \mathbf{c}_r)$ are identical. On the other hand, if we define $f(\mathbf{c}_r, \mathbf{d}_r) = \|\mathbf{d}_r\| - \|\mathbf{c}_r\|$, library-swapping is applicable for $\epsilon > 0$ since we expect $f(\mathbf{c}_r, \mathbf{d}_r) > 0 > f(\mathbf{d}_r, \mathbf{c}_r)$ for most ChIP-enriched regions.

The second condition, $M = \alpha N$, applies a constraint of sequencing depth for ChIP and control, which explains the estimation bias with unbalanced sequencing depth. When $M \neq \alpha N$, the constraint condition can be simply satisfied after a re-sampling procedure by which a subset of reads was randomly retrieved for processing. In detail, if $\alpha N > M$, we retrieve M/α ChIP reads and M control reads; otherwise, N ChIP reads and αN control reads are retrieved.

3.5. The CCAT Framework

In this section, we propose a general-purpose framework, named CCAT (Control based ChIP-seq Analysis Tool), for the significance analysis of ChIP-seq with negative control.

The CCAT framework started with the estimation of noise rate α using the iterative algorithm, followed by a random re-sampling procedure to retrieve a subset of data that satisfies the condition $M = \alpha N$, as described in section 3.3. Next, we exchanged the ChIP and control libraries of the re-sampled data to generate a swapped dataset. Both the re-sampled dataset and the swapped dataset were processed by a region identification module and a region scoring module (see below). A ranking list of candidate ChIP-enriched sites and a list of control sites were generated from the re-sampled dataset and the swapped dataset, respectively. Following Proposition 2, FDR were estimated for the candidate sites using the control sites as reference. The flowchart of CCAT is shown in Figure 3.3.

In our implementation of region identification module, the ChIP-seq reads were first shifted by $l/2$ bps towards their orientations, where l is the average length of DNA fragments. A sliding window was then applied to scan the whole genome. The size of the window was set to be 21 for applications to motif-specific transcription factors, and ranged from hundreds to thousands bps for histone modification study. The shifted ChIP reads and control reads were counted within the window in each step of sliding. We flagged the genomic position at the center of the window if the ChIP read count was more than twice the control read count, corresponding to a minimum expected signal-noise ratio of 1.0. Consecutive flagged positions were merged into non-overlapping candidate regions. Although the ChIP-enriched sites were defined in terms of regions in CCAT, we also applied a peak-finding algorithm [14] to each candidate region for a high-resolution representation of ChIP enrichment site, which facilitates the motif study in some applications.

Figure 3.3: The flowchart of CCAT. Ovals refer to computational steps; rectangles with solid lines refer to the input and output data; rectangles with dash lines refer to the intermediate data.

The region scoring module in CCAT supports multiple options of significance scores, including fold-change, binomial p-value [23][26], Poisson p-value [24], and normalized difference score [41]. In general, these scores can be represented as $f(c, d)$, where c and d take integer values and refer to the total number of reads in a region or around a peak, for control and ChIP libraries respectively. Before counting the reads, the reads were

shifted by half of the average DNA fragment length towards their orientation. Multiple reads mapped to the same location in the orientation were counted only once, in order to reduce the influence of PCR amplification bias. For each significance score, CCAT supports two modes of read counting: “count on region”, by which all reads in a candidate region are counted; and “count on peak”, by which the reads within the sliding window centered by the highest peak of a candidate region are counted. The definitions of the significance scores are as follows:

Fold-change:

$$f(c, d) = \frac{d+\varepsilon}{c+\varepsilon} \quad (3.9)$$

where ε is a pseudo-count to avoid zero-denominator and is set to be one by default.

Binomial p-value:

$$f(c, d) = -\log \left[\sum_{k=d}^{\infty} \frac{(k+c)!}{k!c!} \times (0.5)^{k+c} \right] \quad (3.10)$$

where the Bernoulli probability is 0.5 since the sample size of ChIP and control libraries are already balanced in the re-sampling step of CCAT. We took $-\log()$ operation such that the score is monotonous to the significance.

Normalized difference score:

$$f(c, d) = \frac{d-c}{\sqrt{d+c}} \quad (3.11)$$

Poisson p-value:

$$f(c, d) = -\log \left[\sum_{k=d}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \right] \quad (3.12)$$

where $\lambda = \max(c, 1)$, which is a simplified representation of λ -estimation proposed by Zhang *et al.* [24].

3.6. Spike-in Simulation

To define a standard that is comprehensive enough for the evaluation of computational approaches, we generated spike-in datasets to assess the accuracy of noise rate estimation and FDR estimation. The idea of “spike-in” here is to computationally add simulated ChIP-enriched signal to a control background, and the spike-in loci are referred as standard in the evaluation [41]. Two Nanog ChIP-seq datasets published by different groups (Chen *et al.*, [14]; Marson *et al.*, [43]) were included for generating spike-in ChIP library. These datasets can be treated as biological replicates. We defined the spike-in regions as the binding sites predicted from Marson *et al.*’s dataset, and retrieved spike-in reads for those regions from Chen *et al.*’s dataset. By this, 16,688 spike-in regions were defined, corresponding to 464,757 reads. Due to the variation between replicates, we found reads are not over-represented in some of the spike-in regions, which assembles the false negatives in real datasets. The background reads in the spike-in ChIP library were retrieved from a GFP dataset [14] prepared with non-specific antibody. The noise rate of the spike-in ChIP library was adjustable by merging different proportion of the GFP reads with the spike-in reads. For the control, we generated a WCE (whole-cell extract) dataset, available at http://cmb.gis.a-star.edu.sg/ChIPSeq/tools_.htm. Note that the background reads in the spike-in ChIP and control libraries were retrieved from different sources in order to better simulate the real ChIP-seq experiment, in which the control was prepared with a different antibody or simply input DNA without antibody. All the

libraries used for the spike-in dataset were prepared on mESC (mouse embryonic stem cell), and were mapped by Eland software with identical configurations (26bps, 2 mismatches). In addition to the Nanog data, we also generated H3K4me3 spike-in data in the same manner. 19,631 spike-in regions were defined based on Marson *et al.*'s H3K4me3 dataset [43], and 3,171,004 spike-in reads were retrieved from another dataset published by Mikkelsen *et al.* [21]. A summary of the libraries included for the spike-in data generation is provided in Table 3.1.

Table 3.1: Libraries used for spike-in simulation.

library ID	antibody	# of uniquely mapped reads*	reference
control 1	GFP	3.83M	[14]
control 2	WCE	6.76M	unpublished
Nanog 1	Nanog	6.03M	[43]
Nanog 2	Nanog	8.42M	[14]
H3K4me3 1	H3K4me3	6.94M	[43]
H3K4me3 2	H3K4me3	8.85M	[21]

To assess the accuracy of noise rate estimation, different proportions (10%, 20%, ..., 100%) of GFP library were used as the background for spike-in. Therefore, the real noise rate equals the fraction of GFP reads in the spike-in ChIP library. In Figure 3.4a, the relative estimation error of the iterative algorithm was plotted against the actual noise rate. We found the errors are positive in most of the cases, implying over-estimation of noise rate. The explanation is that some of the background regions determined by Eq. 3.5 contain weak but undetectable signal due to the issue of sequencing depth. Nevertheless, the relative errors are smaller than 5% under the practical sequencing depth, indicating the proposed iterative algorithm reasonably estimated the noise rate. Comparing to the noise rate derived from Ji *et al.*'s method in the software CisGenome [23] (see Section 1.2 for the description of CisGenome and Section 3.3 for its noise rate estimation method), our estimation also achieved better accuracy (Figure 3.4b and 3.4c).

Figure 3.4: Noise rate estimation (to be continued). (a) The relative error of noise rate estimation by proposed iterative algorithm;

(b)

(c)

Figure 3.4: Noise rate estimation (continued). (b,c) Accuracy of the noise rate estimation by the proposed iterative algorithm and CisGenome [23], on (b) Nanog and (c) H3K4me3 spike-in datasets.

Next, we compared the library-swapping based FDR estimation with previous methods. To date, the approaches of FDR estimation for control based ChIP-seq data mainly fall into two categories: (a) correcting binomial p-values using Benjamini-Hochberg (B-H) correction [26][44] or Storey's method [41][42]; (b) estimating expected fraction of false positives by generating empirical background. The library-swapping approach mentioned in Section 3.4 belongs to the latter. Another example of the second category is the eFDR proposed by Nix *et al.* [41]. To calculate eFDR, control library were randomly split in two halves. The eFDR was estimated as the ratio of the number of control enriched regions (control 1 vs. control 2) to the number of experimentally observed enriched regions (ChIP vs. control2).

To assess the accuracy of FDR estimated by different approaches, we employed a spike-in datasets configured as follow: 3 million GFP reads were randomly retrieved and were merged with the Nanog or H3K4me3 spike-in reads to generate ChIP library; 3 million WCE reads were used as control (for eFDR, additional 3 million WCE reads were used as an alternative control). By this configuration, the sequencing depths of ChIP and control were balanced. The whole genome was partitioned into 1 kb non-overlapping regions and the FDR were calculated on the region basis. For a fair comparison, we used binomial p-values as the significance score for all methods. The binomial p-value was proposed by Rozowsky *et al.* to measure the significance of changes between ChIP and control samples based on the read counts in each sample [26]. In their study, Benjamini-Hochberg correction [44] was employed to estimate the FDR from binomial p-value. Prior to calculating p-value, we followed Nix *et al.*'s [41] suggestion to exclude the regions with read counts smaller than ten. This filtering step is necessary for p-value

correction methods, which assume the p-values are uniformly distributed under null hypothesis.

The accuracy of FDR estimation was compared in a dynamic range from 0.001 to 0.2. (Figure 3.5). The plots show the variations and biases between the estimated FDR and real FDR, for different methods. An ideal FDR estimation is expected to show a diagonal line that link the points (0,0) and (1,1) in the space spanned by real FDR and estimated FDR. The results are largely consistent for Nanog (Figure 3.5a) and H3K4me3 (Figure 3.5b) spike-in datasets. We found the empirical approaches significantly outperformed the p-value correction methods, for which the FDR was constantly overestimated. Although both empirical approaches performed well when $FDR > 0.01$, eFDR appeared a non-linear characteristic of estimation bias, and underestimated when $FDR < 0.01$. This may ascribe to slight bias between GFP and WCE libraries prior to random sampling. Such bias could not be modeled using the empirical background that was generated by control-splitting. On the contrary, the library-swapping approach seemed to reasonably compensate this type of bias and achieved satisfactory performance both in accuracy and linearity.

(a)

(b)

Figure 3.5: Comparison of FDR estimation methods using (a) Nanog and (b) H3K4me3 spike-in datasets. The dashed diagonal line depict the ideal approximation.

3.7. Application to Histone Modifications

We applied CCAT to the H3K4me3 (K4) and H3K36me3 (K36) histone modification datasets published by Mikkelsen *et al.* [21]. The GFP library published by Chen *et al.* [14] was used as the negative control. All the libraries included in this study were generated in mouse embryonic stem cells (mESC) based on Illumina-Solexa platform, and the sequenced reads were aligned using Eland with the same configuration (27 bps, 2 mismatches). Therefore, combination of these libraries provides a relevant test-bed for our approach.

In literature, K4 was known to be a promoter marker and K36 has been shown to occupy the gene region as a hallmark of elongation [45]. Our previous study also showed the gene expression is potentially predictable from K4 and K36 marks [29]. Based on a better measurement of ChIP-seq background level and more reliable FDR estimation, the purpose of our application is to study the weak K4 and K36 sites that may not be predictable by previous methods.

In CCAT configuration, we used the normalized difference score [41] as the significance measure. The FDR cut-off was set to be 0.05. The size of sliding window was set to be 500 bps for K4, and 2kb for K36. This setting was determined due to the fact that K4 enriched sites usually appear peak patterns, while K36 signals spread broader regions. As the result, we identified 64,012 K4 regions, corresponding to 75,620 individual peaks. For K36, 19,042 K36 regions were identified.

To assess the sensitivity, we overlapped CCAT identified regions with the UCSC histone modification track (<http://genome.ucsc.edu/>) predicted by Mikkelsen *et al.* [21], as well

as the predictions by PeakSeq [26], with FDR=0.05. As shown in Figure. 3.6a and 3.6b, CCAT predicted 100% more K4 sites and 70% more K36 sites than these two methods did. We further extended our comparison to include MACs [24] and SICER [22]. Again, pair-wise comparison showed that the CCAT prediction is highly sensitive and almost forms a superset of the predicted sites by the other methods, for both K4 and K36 datasets (Figure 3.6c). Readers may refer to Section 1.2 for the descriptions of the previous methods in this comparison.

Figure 3.6: Comparison of CCAT with previous methods (to be continued). (a,b) Venn diagram comparing the predictions by CCAT, UCSC track, and PeakSeq, for (a) H3K4me3 and (b) H3K36me3 dataset;

(c)

Figure 3.6: Comparison of CCAT with previous methods (continued). (c) Pair-wise comparison between CCAT and UCSC track, PeakSeq, MACs and SICER, for H3K4me3 and H3K36me3 datasets.

By traditional arguments, H3K4me3 sites are markers of promoters. Surprisingly, only 26,902 (35.6%) out of 75,620 predicted K4 peaks overlap with known promoters (± 1 kb from TSS), and majority of the promoter-associated K4 peaks are in the top 20k of the ranking list. To understand the chromatin features associated with weaker K4 sites, we ranked the K4 peaks in descending order of intensity (ChIP read counts within 1kb), and compared the list to other two histone modification types: H3K27me3 and H3K4me1 (Figure 3.7c). The ChIP-seq datasets published by Mikkelsen *et al.* [21] and Meissner *et al.* [40] were employed for the comparison. We found H3K27me3 signals are enriched for the K4 peaks ranked from 10k to 20k, corresponding to the K4-K27 bivalent domains that have been proven to be crucial to maintain pluripotency of ESC [30]. Meanwhile, H3K27me3 intensities of the top 10k peaks are relatively low in average. We also observed that H3K4me1, an enhancer marker in mammalian cells [46], is enriched for the peaks ranked after 20k, and is almost depleted for the top-ranking peaks. Therefore, in addition to the traditional arguments on H3K4me3, these lines of evidence suggest distinct chromatin features associated with strong and weak K4 sites. As a hypothesis, the weak K4 signals in the enhancer regions may ascribe to the DNA-loop mediated interactions of enhancers and promoters.

(a)

Figure 3.7: Validation of CCAT predicted histone modification sites (to be continued). (a) Result of qPCR validation on 15 H3K4me3 sites exclusively identified by CCAT, and 10 control sites.

(b)

(c)

Figure 3.7: Validation of CCAT predicted histone modification sites (continued). (b) Comparison of H3K36me3 regions exclusively predicted by CCAT with RefSeq, Ensembl, and MGC gene lists. (c) Distinct chromatin features associated with strong and weak H3K4me3 sites. The histone modification intensities are measured with count per million reads per kb (cpmpkb). The curves were smoothed by a sliding window of length 1000.

3.8. Noise Rates as a Measurement of Data Quality

As mentioned in Section 3.2, the noise rate is a quantitative measurement of ChIP quality in the experiment. In general, a smaller noise rate implies better data quality. Therefore, we asked: is there an empirical threshold of noise rate for determining the quality of a ChIP-seq library? To answer this question, we estimated the noise rates for 22 published ChIP-seq libraries, including 15 transcription factors or transcription regulators [14] and 7 histone modification types [21][40]. As shown in Figure 3.8, the noise rate for most DNA-binding proteins varies from 0.7 to 0.95. One exception is the CTCF library, which results in a noise rate of 0.6, and we can observe 50000+ “clear” peaks in CTCF ChIP-seq profile. For histone modifications, the approximated noise rate varies from 0.3 to 0.9, dependent on the modification types. As a reference, we obtained a noise rate of 0.992 in the comparison between two control libraries, GPF (non-specific antibody) and WCE (whole cell extract). Therefore, the results implicate that a “good” library may have a noise rate below 0.95. However, there could be still detectable peaks from libraries with a noise rate larger than 0.95. Let’s take an extreme case for example. Suppose there is only one binding site in the whole genome, which is highly ChIP-enriched with read count of 10,000 in a 1M sequenced library. The noise rate would be around $990,000/1,000,000 = 0.99$, which is relatively high. Meanwhile, the ChIP quality is good enough to detect this binding site since the read count at the binding site is much higher than the background regions. Therefore, our opinion is that the noise rate is useful to compare the data quality among the experiments of the same design, while the judgment needs to be made in an application-specific manner.

Figure 3.8: Noise rates estimated for published ChIP-seq libraries, including 15 transcription factors or transcription regulators [14] and 7 histone modification types [21][40].

Chapter 4

Identifying Differential Histone Modification Sites

Epigenetic modifications are one of the critical factors to regulate gene expression and genome function. Among different epigenetic modifications, the differential histone modification sites (DHMSs) are of great interest to study the dynamic nature of epigenetic and gene expression regulations among various cell-types, stages or environmental responses. Thus the DHMSs are potentially identifiable by comparing two ChIP-seq libraries. Aiming at identifying DHMSs, we propose an approach called ChIPDiff for the genome-wide comparison of histone modification sites identified by ChIP-seq. Based on the observations of ChIP fragment counts, the proposed approach employs a Hidden Markov Model (HMM) to infer the states of histone modification changes at each genomic location.

4.1. Differential Histone Modification Sites

Eukaryotic DNA is packaged into a chromatin structure consisting of repeating nucleosomes by wrapping DNA around histones. The histones are subject to a large number of posttranslational modifications such as methylation, acetylation, phosphorylation and ubiquitination. The histone modifications are implicated in influencing gene expression and genome function. Considerable evidence suggests several histone methylation types play crucial roles in biological processes [47]. A well-

known example is the repression of development regulators by tri-methylation of histone H3 lysine 27 (H3K27me₃, or K27) in mammalian embryonic stem cell (ESC) to maintain stemness and cell pluripotency [48][49]. Some epigenetic stem cell signature of K27 is also found to be cancer-specific [50]. Moreover, the tri- and di- methylation of H3 lysine 9, are implicated in silencing the tumor suppressor genes in cancer cells [51]. In the light of this, the specific genomic locations with differential intensity of histone modifications, which are called differential histone modification sites (DHMSs) in this paper, are of great interest in the comparative study among various cell-types, stages or environmental response.

ChIP-seq technology has become a robust and comprehensive approach to capture the histone modifications at whole genome scale [12][21]. Inspired by the success of ChIP-seq in identifying histone modification sites in a single library, we asked if the DHMSs could be identified by computationally comparing two ChIP-seq libraries generated from different cell-types or experimental conditions. Mikkelsen *et al.* mapped the H3K4me₃ (K4) and K27 sites in mouse ESC, NPC and embryonic fibroblast (MEF) and compared the occurrence of modification sites in promoter regions across three cell-types [21]. In their comparison, the histone modification sites were first identified for individual libraries; genomic regions marked “with K4/K27” in one cell-type but not in the others were determined as “differential sites”. A limitation of their study is that the modification sites are compared qualitatively but not quantitatively. For some histone modification sites marked “with K4/K27” in all the cell-types, there could exist significant changes of modification levels in quantity. An example demonstrating this limitation is the regulation of *Klf4* by K4, which is known to be positively correlated to gene expression.

The Klf4 promoter was flagged as “with K4” in both ESC and NPC by qualitative analysis hence it could not explain the up-regulation of Klf4 in ESC. On the other hand, quantitative comparison indicated the intensity of K4 in Klf4 promoter is more than 5 fold higher in ESC than in NPC, consistent with the observation of expression change.

To the best of our knowledge, little has been published in literature on the quantitative comparison of two ChIP-seq libraries at genome scale. Triggered by the idea from microarray analysis [52], a simple solution to this problem is to partition the genome into bins and to compute the fold-change of the number of ChIP fragments in each bin. However, fold-change approach is sensitive to the technical variation caused by random sampling of ChIP fragments. In this paper, we propose an approach called ChIPDiff to improve the fold-change approach by taking into account the correlation between consecutive bins. We modeled the correlation in a hidden Markov model (HMM) [53], in which the transmission probabilities were automatically trained in an unsupervised manner, followed by the inference of the states of histone modification changes using the trained HMM parameters.

4.2. Determining Putative Histone Modification Sites

Given two ChIP-seq libraries, L_1 and L_2 , the first step for identifying DHMSs is to determine the putative sites that involve histone modifications either in L_1 or L_2 . This section details this step.

Tags in the raw data generated from a ChIP-seq experiment were mapped onto the genome to obtain their positions and orientations. Due to the PCR process in ChIP-seq experiments, multiple tags may be derived from a single ChIP fragment. To remove the redundancy, tags mapped to the same position with the same orientation were treated as a single copy. Note that in ChIP-seq protocol a tag is retrieved by sequencing one end of the ChIP fragment, of which the median length is around 200bps [12][18]. Hence we approximated the center of corresponding ChIP fragment by shifting the tag position by 100bps towards its orientation. The whole genome was partitioned into 1k-bps bins and the number of centers of ChIP fragments were counted in each bin.

After the above pre-processing procedures, a profile of ChIP fragment counts was generated. Considering a genome with m bins, the profiles of L_1 and L_2 are represented as $X_1 = \{x_{1,1}, x_{1,2}, \dots, x_{1,m}\}$ and $X_2 = \{x_{2,1}, x_{2,2}, \dots, x_{2,m}\}$, respectively, where $x_{i,j}$ is the fragment count at the j th bin in L_i . To depict the combined enrichment of fragments in each bin, we defined a score F normalized against the sequencing depth:

$$F(i) = \frac{x_{1,i}}{n_1} + \frac{x_{2,i}}{n_2}, \quad i = 1, 2, \dots, m \quad (4.1)$$

where n_1 and n_2 are the total number of sequenced fragments in L_1 and L_2 , i.e. $n_1 = \sum_i x_{1,i}$,

$$n_2 = \sum_i x_{2,i}.$$

Mikkelsen *et al.* and Robertson *et al.* pointed out that not all the bins can be interrogated in the tag mapping procedure, mainly due to the existence of repeat region [18][21]. Let

η denote the fraction of “valid” bins in the genome, the expectation of the score F at a “valid” bin is $\sum_i F(i) / (m \times \eta)$, which equals $2 / (m \times \eta)$. Mikkelsen *et al.* estimated $\eta \approx 0.7$ for mouse genome. If a bin has an F score larger than $2 / (m \times \eta)$, we flagged it as a putative histone modification site. Consecutive modification sites within 1k bps apart from each other were merged into histone modification regions.

4.3. Quantitative Comparison of Intensity by Fold-change

For the convenience of denotation and description, the methods in the rest of the chapter will be introduced based on the putative modification regions defined in Section 4.2. Considering a region that consists of k bins, we use the notation $x_{1,i}$, $x_{2,i}$ for the ChIP fragment counts in L_1 and L_2 , respectively, at the i th bin in that region ($i = 1, 2, \dots, k$).

Histone modifications exhibit a variety of kinetics and stoichiometries [54]. For a ChIP-seq experiment, we define the modification intensity at the i th bin in library L_j to be the probability of an arbitrary ChIP fragment captured from the i th bin in the ChIP process, denoted $p_{j,i}$. Since the extraction and sequencing of ChIP fragments is a random sampling process, the posterior probability of observing $x_{j,i}$ fragments at the i th bin in library L_j , conditional on the intensity $p_{j,i}$, follows a binomial distribution:

$$\Pr(x_{j,i} | p_{j,i}) = \binom{n_j}{x_{j,i}} p_{j,i}^{x_{j,i}} (1 - p_{j,i})^{n_j - x_{j,i}} \quad (4.2)$$

where n_j is the total number of reads in library L_j . Note that the binomial model proposed here is approximately equivalent to the Poisson model proposed in Section 3.2. Readers may refer to Page 49 for the rationale.

We further assume the prior probability of $p_{j,i}$ follows a beta distribution:

$$\Pr(p_{j,i}) = \frac{1}{B(\alpha, \beta)} p_{j,i}^{\alpha-1} (1 - p_{j,i})^{\beta-1} \quad (4.3)$$

where $B(\alpha, \beta)$ is the beta function.

The beta distribution $Beta(\alpha, \beta)$ has the following features [55]:

- a) It is the conjugate prior of binomial distribution [56]. Specifically, suppose $\Pr(x | p)$ follows binomial distribution $Binom(x; n, p)$, where p is the Bernoulli probability, n is the total number of samples, and x is the binomial variable; if the prior probability $\Pr(p)$ follows a beta distribution $Beta(\alpha, \beta)$, then $\Pr(p | x)$ follows beta distribution $Beta(\alpha + x, \beta + n - x)$;
- b) The expectation of $Beta(\alpha, \beta)$ is $\alpha / (\alpha + \beta)$;
- c) In assuming a prior beta distribution $Beta(\alpha, \beta)$, taking larger values of α and β suggests stronger prior belief.

Based on the features a) and b), we can derive that the conditional probability $\Pr(p_{j,i} | x_{j,i})$ follows a beta distribution $Beta(\alpha + x_{j,i}, \beta + n_j - x_{j,i})$, of which the expectation is $(\alpha + x_{j,i}) / (\alpha + \beta + n_j)$. In our application, we set the parameters α and β

to be 1 and m , where m is the total number of bins. By this setting, the prior expectation of $p_{j,i}$ for each bin is $1/m$.

We define a DHMS as a bin in which the ratio of intensities between L_1 and L_2 is larger than τ (L_1 enriched DHMS) or smaller than $1/\tau$ (L_2 enriched DHMS), where τ is a pre-determined threshold, and $\tau \geq 1.0$. A simple solution for identifying DHMSs is to estimate the fold-change of expected intensity (preferably in term of log-ratio) from the ChIP fragment counts, as follow:

$$\log \left[\frac{E(p_{1,i} | x_{1,i})}{E(p_{2,i} | x_{2,i})} \right] = \log \left[\frac{(\alpha + x_{1,i})(\alpha + \beta + n_2)}{(\alpha + x_{2,i})(\alpha + \beta + n_1)} \right] \quad (4.4)$$

An example of the log-ratio estimation based on Eq. 4.4 is shown in Figure 4.1a.

A drawback of the fold-change approach is that it is prone to the technical variation caused by random sampling. Figure 4.1b shows an RI-plot [52] to depict the variation of the log-ratio dependent on the intensity. When the intensity is relatively small, the variation of log-ratio becomes too high, which may result in considerable false positives.

(a)

(b)

Figure 4.1: (a) An example of the log-ratio estimation of H3K27me3 intensity between mouse ESC and NPC. Bin size set to be 1k; displayed genomic region range from chr14:117,100,000 to 117,130,000; data retrieved from Mikkelsen *et al.*'s dataset [21]; (b) An RI-plot for chromosome 19 in K27 data.

4.4. An HMM-based Approach for Identifying DHMSs

Histone modifications usually occur in continuous regions that span a few to hundreds or even thousands of nucleotides. Hence one may expect strong correlation between consecutive bins in the measurements of intensity changes. This argument is supported by our observations from ChIP-seq profile. As an example, the log-ratio profile in Figure 4.1a has an autocorrelation of 0.84. In ChIP-chip data analysis, Li *et al.* have designed an HMM to model the correlation of signals between consecutive probes and successfully applied it for the identification of p53 binding site [57], suggesting the potential ability of HMM for identifying DHMSs in our study. Here we propose a HMM-based approach called ChIPDiff to solve the problem.

Let s_i be the state of histone modification change at the i th bin ($i = 1, 2, \dots, k$). Based on the definition of DHMS in section 4.3, the state s_i takes one of the following three values:

- α_0 : non-differential site, if $1/\tau \leq p_{1,i} / p_{2,i} \leq \tau$;
- α_1 : L_1 enriched DHMS, if $p_{1,i} / p_{2,i} > \tau$;
- α_2 : L_2 enriched DHMS, if $p_{1,i} / p_{2,i} < 1/\tau$.

We modeled the inter-bin correlation as a first-order Markov chain such that $\Pr(s_i | s_0, s_1, \dots, s_{i-1}) = \Pr(s_i | s_{i-1})$, where s_0 is the start state before the first bin in the region. A HMM was implemented to infer the posterior probability distribution of the states from the observations of fragment counts. The HMM is characterized by three

Figure 4.2: The graphic representation of the HMM used in ChIPDiff.

features: the prior probability of the start state s_0 , the emission probability $\Pr(x_{1,i}, x_{2,i} | s_i)$, and the transmission probability $\Pr(s_i | s_{i-1})$. The graphic representation of the HMM is shown in Figure 4.2.

The initial state s_0 is fixed to take the value α_0 since we assume the region starts from the genomic locations where histone modification is depleted in both libraries.

We derived the emission probability $\Pr(x_{1,i}, x_{2,i} | s_i)$ by integrating $p_{1,i}$ and $p_{2,i}$ over all possible values constrained by s_i :

$$\Pr(x_{1,i}, x_{2,i} | s_i) = \frac{\Pr(x_{1,i}, x_{2,i}, s_i)}{\Pr(s_i)} = \frac{\iint \Pr(x_{1,i}, x_{2,i}, s_i | p_{1,i}, p_{2,i}) \Pr(p_{1,i}, p_{2,i}) dp_{1,i} dp_{2,i}}{\iint \Pr(s_i | p_{1,i}, p_{2,i}) \Pr(p_{1,i}, p_{2,i}) dp_{1,i} dp_{2,i}} \quad (4.5)$$

Note that, if $p_{1,i}$ and $p_{2,i}$ satisfy the condition constrained by s_i , then s_i is always true, otherwise always false. The above equation can be re-written as:

$$\Pr(x_{1,i}, x_{2,i} | s_i) = \frac{\iint \Pr(x_{1,i}, x_{2,i} | p_{1,i}, p_{2,i}) \Pr(p_{1,i}, p_{2,i}) dp_{1,i} dp_{2,i}}{\iint_{s_i} \Pr(p_{1,i}, p_{2,i}) dp_{1,i} dp_{2,i}} \quad (4.6)$$

Since $p_{1,i}$ and $p_{2,i}$ are independent variables, $x_{1,i}$ only depends on $p_{1,i}$, and $x_{2,i}$ only depends on $p_{2,i}$, we have:

$$\Pr(x_{1,i}, x_{2,i} | s_i) = \frac{\iint \Pr(x_{1,i} | p_{1,i}) \Pr(x_{2,i} | p_{2,i}) \Pr(p_{1,i}) \Pr(p_{2,i}) dp_{1,i} dp_{2,i}}{\iint_{s_i} \Pr(p_{1,i}) \Pr(p_{2,i}) dp_{1,i} dp_{2,i}} \quad (4.7)$$

where $\Pr(x_{j,i} | p_{j,i})$ follows binomial distribution defined in Eq. 4.2 and $\Pr(p_{j,i})$ follows beta distribution defined in Eq. 4.3.

The transmission probability table was trained using the Baum-Welch algorithm [58], which takes expectation maximization (EM) steps to iteratively estimate the parameters of the HMM from hidden states in an unsupervised manner. In the training process, the transmission parameters were initialized to be uniform and the start state s_0 and the emission probabilities were fixed as described above. Since the transmission probability table is identical in the whole genome, it was trained by cumulating the transmission frequencies in all the putative histone modification regions.

In the last step of ChIPDiff, the probability distributions of the states in each bin were inferred using forward-backward algorithm. Bins with posterior probability larger than a confidence threshold ρ ($0 < \rho < 1$) for $s_i = \alpha_1$ or $s_i = \alpha_2$ were identified as DHMSs. Consecutive DHMSs with no gap between them were merged into DHMS regions.

The most computationally expensive step in ChIPDiff is the training of the transmission probability table. Two strategies were employed to reduce the computational cost: (a) the integrals for emission probabilities were numerical computed and were compiled into a look-up table, prior to the training of the HMM; (b) we allowed the transmission probability table to be trained based on a subset randomly selected from the putative histone modification regions.

4.5. Evaluation with H3K27me3 Data

We applied ChIPDiff to Mikkelsen *et al.*'s dataset publicly available at http://www.broad.mit.edu/seq_platform/chip/. The performance of ChIPDiff was evaluated by comparing the H3K27me3 (K27) libraries between mouse ESC and NPC. H3K27me3 was selected for the evaluation since its DHMSs in highly conserved noncoding elements (HCNEs) have been implicated in literature [49]. Moreover, K27 preferentially marks gene region and functions as a repressor, which facilitated our indirect validation using expression data. We compared the K27 ESC library and NPC library with ChIPDiff, in which the fold-change threshold τ was set to be 3.0 and the confidence threshold ρ was set to be 0.95. The HMM was trained with 10,000 randomly selected histone modification regions. 26,230 bins were identified to be DHMSs,

corresponding to 4,722 continuous regions. Among them, 3,833 (81.2%) regions are ESC enriched and 889 (18.8%) are NPC enriched, implying a global trend of K27 depletion upon cell differentiation.

We first assessed the capability of ChIPDiff in identifying the biologically significant DHMSs, i.e., sensitivity. Bernstein *et al.* have shown that K27 is enriched in highly conserved noncoding elements (HCNEs) in ESC, repressing a number of development regulators to maintain the stemness of the cell [49]. These histone marks are depleted in diverse differentiated cells. From HCNEs, we selected 223 genes of which the expressions were studied by Mikkelsen *et al.* [21]. Since K27 functions as a gene repressor, we reasoned that some of those HCNE genes marked by K27 will be up-regulated in NPC and DHMSs should be identified at these genes. As expected, a subset containing 30 genes were determined to be up-regulated with the criterion of more than 4-fold. Among them, 24 (80%) are marked by DHMSs identified by ChIPDiff in promoter region $\pm 1\text{kb}$ from transcription start site (TSS). In contrast, only 37 (19.2%) out of the 193 genes that are not up-regulated in NPC are marked by DHMSs.

To test the specificity of ChIPDiff result, we need to estimate the fraction of falsely identified DHMS regions that are not cell-specific. For this purpose, we partitioned each library into two technical replicates: $L_{ESC,rep1}$ and $L_{ESC,rep2}$ for ESC, $L_{NPC,rep1}$ and $L_{NPC,rep2}$ for NPC. The replicates consist of tags retrieved from different lanes in ChIP-seq experiments, with similar sequencing depth (see Table 4.1 for the assignment of lanes to replicates). We generated two new libraries by merging the tags in $L_{ESC,rep1}$ and $L_{NPC,rep1}$, $L_{ESC,rep2}$ and $L_{NPC,rep2}$, respectively. Since the replicates are of similar sequencing depth,

the difference between these two libraries should not be cell-specific and only reflect the technical variations in the experiments. Comparing these non-cell-specific controls, 9 differential regions were identified by ChIPDiff. Hence, we approximated a false positive rate of 0.19% (9/4722) for the DHMS regions identified in cell-specific comparison.

We also tested the reproducibility by conducting two independent passes of cell-specific comparison: $L_{ESC,rep1}$ vs. $L_{NPC,rep1}$, and $L_{ESC,rep2}$ vs. $L_{NPC,rep2}$. We first measured the statistical significance of overlap between two passes based on hyper-geometric distribution $hyper(k;a,b,n)$, where k is the number of bins identified as DHMSs in both passes, a and b are the number of DHMSs determined in individual passes, and n is the total number of bins in the entire genome. We found the p-values is smaller than 1E-308, indicating the results are reproducible from a statistical point of view. To better depict the significance of overlap, we defined a reproducibility score as the ratio of the number of DHMSs identified in both passes, to the average number of DHMSs in individual pass. The reproducibility score is analogous to the fraction of overlap in the Venn diagram. As the result, we obtained a reproducibility score of 57.4% for ChIPDiff. Note that the reproducibility is conditional on the sequencing depth of the replicates, which ranges from 3 to 4 million tags in our assessment (Table 4.1).

To compare the performance among different methods, we repeated the sensitivity, specificity and reproducibility tests for fold-change and qualitative method. In qualitative method, K27 modification sites were identified for ESC and NPC individually using the binning approach proposed by Mikkelsen *et al.* [21], and bins marked as K27 site in only one cell-type were identified to be DHMSs. Consecutive DHMSs were merged into

DHMS regions as well. For a fair comparison, the thresholds were adjusted to allow similar number of DHMS regions to be identified for all three methods (the numbers are not identical because the thresholds take discrete values). The evaluation results are summarized in Table 4.2. ChIPDiff outperformed the other two methods in all three aspects. Fold-change approach and qualitative method had much higher false positive rates, indicating these methods are sensitive to technical variation and experimental bias.

Table 4.1: Assignment of lanes in Mikkelsen et al.'s ChIP-seq data to the replicates defined in the paper

replicate id	flowcell	Lane	#tags	cell type
L _{ESC} , rep1	2491	2	337495	ESC
L _{ESC} , rep1	2491	3	1403845	ESC
L _{ESC} , rep1	2491	4	340808	ESC
L _{ESC} , rep1	2491	5	1361536	ESC
L _{ESC} , rep2	2491	6	242432	ESC
L _{ESC} , rep2	2491	7	1336598	ESC
L _{ESC} , rep2	3481	6	1515212	ESC
L _{NPC} , rep1	3099	6	1481543	NPC
L _{NPC} , rep1	3099	7	1827307	NPC
L _{NPC} , rep1	3484	5	973020	NPC
L _{NPC} , rep2	3484	4	3667936	NPC

Table 4.2: Comparison of the performance of ChIPDiff, fold-change approach and qualitative method, based on H3K27me3 data. FPR refers to false positive rate.

	ChIPDiff	fold-change	qualitative method
# of DHMS regions in cell-specific comparison	4,722	4,958	4,790
FPR estimated from non-cell-specific control (%)	0.19	10.8	52.3
Detection rate on HCNE DHMSs (%)	80.0	63.3	73.3
Reproducibility score (%)	57.4	23.4	43.8

In practice, a binning strategy may fail to identify a modification site right on the junction of two bins. Conventionally, people solve this problem by assigning bins such that consecutive bins overlap with each other by half of their length. However, the observation of tag counts in consecutive bins would not be independent in this case, which results in theoretical problem in the subsequent HMM inference in ChIPDiff. An alternative approach is to run ChIPDiff for two passes. In the second pass, the phase of binning shifts by half of the bin size in relative to the first pass. By merging the results in both passes, one may improve the sensitivity of the method.

On the other hand, since the HMM approach takes into account the correlation between consecutive bins, ChIPDiff is theoretically less sensitive to the partitioning problem. To test this argument, we performed ChIPDiff on K27 data for two runs of which the phases of binning shift by 500 bps. We found nearly 90% of the DHMSs are identified in both runs even if the phases are different, as shown in the Venn diagram in Figure 4.3a. For comparison, we tested with the fold-change approach and found only around 50% of the

DHMSs in overlap (Figure 4.3b). This experiment confirmed that ChIPDiff is less sensitive to the partitioning problem while this problem really matters with fold-change approach.

Figure 4.3: Venn diagram demonstrating the repeatability of identified DHMSs for (a) ChIPDiff and (b) fold-change, when the phase of binning changes. Bin size was set to be 1k bps; phase 2 shifts by 500 bps in relative to phase 1.

4.6. Application to H3K4me3 and H3K36me3 Data

We extended our study to trimethylations on K4 and K36. Both histone modification types positively regulate gene expression but in different manner. Guenther *et al.* revealed K4 marks the active promoters where the transcription of the genes is initiated, while K36 occupies the gene region as a hallmark of elongation [45]. Our previous study also showed that K4, together with K27, establishes distinct genomic domains of active and inactive chromatin structures in human ESC [30]. Thus it attracted our interest to study the DHMSs of these histone modifications between ESC and NPC. Moreover, K4 sites usually appear in punctated pattern sharply around transcription start site in ChIP-seq profile, while K36 sites appear in a much broader pattern, providing a comprehensive test-bed for evaluating the adaptability of our approach to diverse histone modification types.

We processed the libraries with the same ChIPDiff configurations as mentioned in Section 4.5. The results are summarized in Table 4.3. Consecutive bins identified as DHMSs were merged into regions. Strikingly, the number of ESC enriched K4 DHMSs is much larger than NPC enriched ones. Considering such imbalance was also observed for K27, we hypothesized it may be associated with the bivalent chromatin structure marked by K4 and K27 [49]. In further analysis, we found 1,961 (51.2%), out of 3,833 ESC enriched K27 DHMS regions overlap with ESC enriched K4 DHMSs. In contrast, K36 and K27 seemed to be mutually exclusive: only 8 (0.21%) of these 3,833 regions overlap with ESC enriched K36 DHMSs.

Table 4.3: A summary of DHMSs identified from H3K4me3 and H3K36me3 libraries.

	H3K4me3		H3K36me3	
	ESC enriched	NPC enriched	ESC enriched	NPC enriched
# of DHMS bins	32,384	3,742	15,111	16,719
# of DHMS regions	12,976	1,768	1,158	1,228
# of RefSeq genes marked	3,877	211	747	417

To study the genomic patterns of DHMS regions, we plotted the length distributions of DHMS regions for K27, K4 and K36, individually, as shown in Figure 4.4. Approximately 70% of K4 DHMS regions are of length below 2k bps, and more than 90% below 5k bps. This observation implies punctated pattern of the K4 DHMS regions. In contrast, K36 DHMS regions are much broader: nearly 70% of them are more than 5k bps and 40% above 10k bps. K27 appears a mixture of punctated and broad pattern. Less than 10% of the K27 and K36 DHMS regions consist of a single 1k-bp bin, indicating the resolution we used in this study is sufficient for these two modification types. For K4, a higher resolution may be helpful.

Next we study the correlation between DHMSs and gene expression. We annotated the RefSeq genes with DHMS regions and expression data published by Mikkelsen *et al.* RefSeq genes were retrieved from UCSC database. To remove the redundancy, the longest ORF was selected for gene annotation if multiple transcripts are mapped to the same gene, which resulted in 18,795 unique genes in total. As shown in Figure 4.5, K4

and K36 co-regulate the gene expression with strong significance. This observation is consistent with the conclusion by Guenther *et al.* [45]. Among 1,085 genes up-regulated in ESC, 791 (72.9%) are associated with ESC enriched K4 or K36 DHMSs, suggesting the gene expression is potentially predictable from DHMSs. Notably, two key transcription factors in ESC, Nanog and Oct4, are marked by DHMSs of both K4 and K36, implying the critical roles played by these histone modification marks in ESC by interfering the transcription regulatory network.

Figure 4.4: Length distribution of DHMS regions for K4, K27, and K36.

Figure 4.5: Combinatorial effect of H3K4me3 and H3K36me3 on gene expression between ESC and NPC. Up/Down regulations were determined by the criterion of 4-fold change.

Chapter 5

Reconstruction of Transcriptional Regulatory Network

In this chapter, we propose the computational approaches for the analysis on a mESC ChIP-seq dataset [14] that involves multiple transcription factors (TFs). The goal of our analysis is to reconstruct the ES-specific transcriptional regulatory network. The details of the dataset were described in Table 2.2. Based on single library analysis, we identified hundreds to tens of thousands of binding sites for these TFs. However, an integrative analysis of these ChIP-seq libraries is necessary for the reconstruction of the network. Here we focus our study on four aspects, step-by-step:

- *De novo* motif discovery;
- Multiple TF co-localization;
- Correlation of binding events and gene expression;
- Reconstruction of transcriptional regulatory network.

5.1. *De Novo* Motif Discovery

In genetics, a motif refers to a sequence pattern that is widespread and has, or is conjectured to have, a biological significance. In transcription regulation, the definition of a motif is specified as a short nucleotide sequence pattern that is frequently observed at

the TF binding site. Studying the motif pattern helps researchers to gain a better understanding of transcription regulation from a molecular perspective, thus is important in the reconstruction of transcription regulatory network. For example, through motif analysis on the Oct4, Sox2, and Nanog TFs, scientists have revealed distinct regulation mechanisms that depend on the order of motif arrangement [59].

Two of the major characteristics of ChIP-seq are the high specificity and the high-resolution representation of binding sites. Taking advantage of these characteristics, *de novo* motifs could be easily found by searching the consensus DNA sequences among the proximal regions of hundreds to tens of thousands predicted peaks. To date, motif discovery is becoming one of the routine procedures in the post-analysis of ChIP-seq data designed for TF study.

To determine the DNA-binding motifs of the TFs under study, sequences ± 100 bps from the top 500 binding peaks with highest intensities were selected for each factor. Genomic repeats were masked since the redundant sequences in repeat regions will result in significant false positives in motif discovery. The program Weeder [60] was used to find overrepresented consensus sequences.

Weeder reports an $N \times 4$ alignment matrix, $F = \{f_{ib}\}$, for the motif, where N is the motif length. The alignment matrix shows the frequency, f_{ib} , of each nucleotide, where $b \in \{A, C, G, T\}$ at each position $i = 1, 2, \dots, N$ in the motif. It represents the nucleotide conservation at each position in the known set of motif occurrences.

For motif scanning we first transform the alignment matrix, $F = \{f_{ib}\}$, into a positional weight matrix (PWM), $W = \{w_{ib}\}$ using the formula

$$w_{ib} = \ln \left(\frac{\bar{f}_{ib}}{f'_b} \right) \quad (5.1)$$

where $\bar{f}_{ib} = \frac{f_{ib}}{\sum_b f_{ib}}$ is the normalized frequency of nucleotide b at position i in the motif,

and f'_b is the normalized frequency of nucleotide b in background sequences. The weight matrix gives the log likelihoods for each of the nucleotides at each position of the motif.

For a given nucleotide sequence, $S = s_1 s_2 s_3 \dots s_N$, where $s_i \in \{A, C, G, T\}$ and N is the sequence length, the match score with the PWM is computed as:

$$score = \frac{1}{N} \sum_{i=1}^N w_{is_i} \quad (5.2)$$

The above score formula takes a summation of the PWM weights. Therefore it is a sum of log-likelihoods, which corresponds to a product of the likelihoods at individual base positions.

Because of the high resolution in defining the binding sites offered by the high sequence depth coverage, overrepresented motifs could be uncovered from 12 of the 13 sequence-specific transcription factors (excluding E2f1) (Figure 5.1). Among them, the Tcfcp2l1 motif has not been reported previously, and the rest eleven motif matrices are very close to the corresponding consensus sequences in literature or public motif dataset. Interestingly, we found the de novo motifs of the key ESC-specific regulators Nanog,

Oct4, Sox2, and the signaling component Smad1, are highly similar and resemble the Oct4-Sox2 co-motif identified in our previous study [21]. This observation suggests that these factors may be co-localized to regulate mESC. To validate this hypothesis, and to find out other TFs that may involve in the co-localization, systematic computational analysis on the TF co-localization is necessary.

Figure 5.1: De novo motif matrices discovered from ChIP-seq peaks.

5.2. Co-localization of Multiple Transcription Factors

Upon close examination of the binding profiles from these thirteen TFs, we found that a subset of binding sites were bound by more than one TFs (two examples are shown in Figure 5.2). While some TF-bound rich regions would be real, many might appear due to random chance alone. To distinguish regions enriched beyond random expectation from noise, we developed an algorithm that takes into account both the abundance of and the signal intensities of the bound regions for each transcription factor.

Figure 5.2: Examples of co-localization of multiple transcription factors.

The co-localization analysis started with a TFBS clustering algorithm. 189,225 predicted binding peaks from 13 TF libraries were included as input for clustering. Consecutive binding peaks within 100 bps are iteratively grouped into co-localization loci. Each locus was assigned a co-localization score, defined as the number of TFs with binding peak(s) at that locus. To check the statistical significance of co-localization, we randomly assigned the peaks to the genome, and performed the clustering procedure on the randomized peaks. The histogram of co-localization scores from predicted peaks was plotted against that from random simulation. As shown in Figure 5.3, Loci bound by 4 or more TFs are highly significant ($\text{FDR} < 0.001$), and there is a total of 3583 such Loci, named Multiple TF-binding Loci (MTL). Therefore, the result of the clustering method indicates that the co-localization of TFs is statistically significant at the predicted MTLs.

To further dissect the composition of the MTLs, we examined the co-occupancy of different factors found in the 3583 MTLs. We rendered a TF-loci occupancy matrix, in which a Boolean value in the i th row and the j th column represents the availability of the j th TF at the i th MTL. From this matrix, we compute the Pearson correlation between each pair of columns to measure the co-occupancy of two TFs. The Pearson correlations were then used to group the highly co-occupied TFs based on hierarchical clustering (Figure 5.4). Among the 13 TFs, Nanog, Sox2, Oct4, Smad1, and STAT3 tend to co-occur quite often, as do members of a second, distinct group comprised of n-Myc, c-Myc, E2f1, and Zfx. Strikingly, we found more than 70% of the p300 binding sites ($p < 1e-300$) are in the MTLs with Nanog, Sox2, Oct4, Smad1 or STAT3. On the other hand, MTLs

with n-Myc, c-Myc, E2f1, or Zfx are not enriched by p300 peaks, but mostly occur in the promoter of genes.

From a perspective of biological study, our results indicate that Smad1 and STAT3 share many common target sites with Nanog, Oct4, and Sox2, and reflect a point of convergence of the two key signaling pathways (via Smad1 and STAT3) with the core circuitry defined by Nanog, Oct4, and Sox2. The co-localization of these factors is also associated with the co-factor p300, which is known to be one of the markers of enhancers [46]. Consequently, the MTLs bound by these factors define a number of genomic “hotspots” that regulate the gene expression in mESC.

Figure 5.3: Histogram of the number of TFs bound per co-bound locus. The distribution of randomly occurring co-bound loci is obtained by simulation.

Figure 5.4: Co-occurrence of TF groups within MTLs. Colors in the heat map reflect the correlation of each pair of TFs in MTLs.

5.3. Correlation of Binding Events and Gene Expression

Next we sought to establish the correlation between TF binding events and gene expression. In previous studies of ChIP data, a commonly employed approach for determining target genes of a TF is to associate the genes with TF binding sites based on their relative distance. For example, Loh *et al.* identified the genes with TF binding sites within 50kb from the transcription units as target genes of Oct4 and Nanog [59]. However, the threshold for the distance-based criterion could be different for variant TFs:

the TFs enriched in promoters may require a shorter distance of threshold than the TFs enriched in enhancers. In our method, we modified the distance-based approach by assigning an association score for each pair of gene and TF, where the score takes a value between 0 and 1. A higher score implies higher chance of the gene being the target of the TF. By this way, we avoided arbitrary setting of the cut-off threshold.

Letting g represent a non-redundant gene and tf represent a TF in our dataset, we assume the probability of the gene g being the target of tf is dependent on the location of the nearest tf binding site, denoted by $l(g,tf)$, in relative to the TSS of g . To quantitatively estimate the location-dependent association between genes and TFs, we first aligned the TSSs of all 17,442 genes and divided the genomic locations into 16 bins, separated by $\{-100k \text{ bps}, -50k \text{ bps}, -20k \text{ bps}, -10k \text{ bps}, -5k \text{ bps}, -2k \text{ bps}, -1k \text{ bps}, 0 \text{ bps}, 1k \text{ bps}, 2k \text{ bps}, 5k \text{ bps}, 10k \text{ bps}, 20k \text{ bps}, 50k \text{ bps}, 100k \text{ bps}\}$, and the histogram of $l(g,tf)$ was calculated based on bins. Next, we randomized the tf binding sites by uniformly placing them in the whole genome and calculated the histogram of $l(g,tf)$ based on the random model. Comparing these two histograms, we are able to approximate the proportion of non-random tf -gene association for each bin. Given a pair, g and tf , letting k represent the index of bin corresponding to $l(g,tf)$, the association score was calculated by:

$$Score(g,tf) = \begin{cases} 0, & \text{if } Hist_{real}(k) < Hist_{rand}(k) \\ \frac{Hist_{real}(k) - Hist_{rand}(k)}{Hist_{real}(k)}, & \text{if } Hist_{real}(k) \geq Hist_{rand}(k) \end{cases} \quad (5.3)$$

where $Hist_{real}(k)$ represents the histogram calculated from the real data and $Hist_{rand}(k)$ represents the histogram calculated from random model.

Figure 5.5: TF-gene association scores for 13 TFs and Suz12.

The TF-gene association scores are plotted for the TFs in Figure 5.5. The scores are peaked around TSS for all the TFs, but the widths of the peaks vary from nearly 100k bp (Nanog, Oct4, Sox2, Smad1, STAT3) to 2k bps (E2f1, Suz12), indicating different concentration of binding sites in relative to TSS.

By assigning an association score to each pair of TF and gene, we generate an association matrix M , where each row corresponds to a RefSeq gene and each column corresponds to a TF. We performed k-mean clustering on M to group the genes with similar TF associations [61]. Prior to clustering, each column was normalized by dividing the average score of the column and was mean-centered. We used the software Cluster 3.0 for the clustering, which allows multiple trials to find the best clustering solution (<http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm>).

As the result, we grouped the genes into five classes (Figure 5.6a). Class I genes are enriched in binding sites for Nanog, Oct4, Sox2, Smad1 and STAT3. Class II genes are bound heavily by c-Myc and n-Myc. Class III genes show enrichment (more than 1 fold) in binding by n-Myc, Klf4, Esrrb, Tcfcp2l1, Zfx and E2f1. Class IV is highly enriched in Suz12 bound genes while class V genes are deficient in all the transcription factors. In total, 48% of genes are deficient in transcription factor binding by the thirteen transcription factors (class IV and class V). We note that E2f1 and Suz12 localization is essentially mutually exclusive, suggesting that polycomb repressor complexes inhibit the binding of E2f1 to its target sites.

To establish the correlation between different classes of genes and gene expression, we analyzed the expression level of genes found in each group using published microarray

dataset for undifferentiated ES cells [62]. Overall, the expression level is highest for genes in class II followed by class I and class III ($p < 1E-20$ as determined by Kolmogorov-Smirnov test) (Figure 5.6b). We found genes in class IV and class V are either not expressed or expressed at a very low level in ES cells. More generally, the analysis shows that genes that have multiple bound transcription factors (classes I to III) are typically expressed at higher levels in ES cells.

To further characterize the gene expression profiles of each class, we used a microarray dataset that interrogated the transcriptome dynamics of retinoic acid (RA)-induced differentiation [62]. Based on 2-fold change, the genes in this dataset were divided into three categories. They are genes up-regulated in ES cells, non-differentially expressed genes and genes down-regulated in ES cells. As shown in Figure 5.6c, Class I genes constitute less than 10% of the non-differentially expressed genes and genes down-regulated in ES cells. This compares to 24% of the genes in the up-regulated category. The percentage of class II genes is only 12% among non-differentially expressed genes, but 36% in the up-regulated set. Hence, class I and class II genes are 2.7 ($p = 8.14E-52$) and 2.9 ($p = 1.28E-91$) fold enriched, respectively, in genes up-regulated in ES cells. In contrast, class IV and class V genes are underrepresented in this set. Class III is slightly enriched in genes that are down-regulated in ES cells, but not enriched in genes that are preferentially up-regulated in ES cells. In summary, our global analysis showed that 60% of genes up-regulated in ES cells are from class I and class II. Most importantly, the result demonstrates that gene clustering based on TF occupancies has the potential to predict ES cell-specific gene expression. This suggests that the TF binding patterns of these two groups are relevant in specifying ES cell-specific expression.

(a)

Figure 5.6: Association of TF binding and gene expression (to be continued). (a) Five classes of genes grouped based on TF-gene association score.

(b)

(c)

Figure 5.6: Association of TF binding and gene expression (continued). (b) Histogram of the gene expression for genes in five classes; (c) Proportion of different classes of genes found in differentially and non-differentially expressed genes in a published expression dataset[62].

5.4. The Core Transcriptional Regulatory Network in mESC

Traditionally, people reconstruct the transcriptional regulatory network based on the co-expressions of genes, which could be measured in microarray experiments. However, it is difficult to predict the direct targets of TFs using microarray-based approach, since the co-expression of two genes may involve multiple-stage transcriptional regulation in cascade. With the TF binding information rendered from ChIP-seq data, we targeted at the reconstruction of a more accurate transcription regulatory network with detailed description of direct regulatory relationship.

The self-renewing state of undifferentiated ES cells is characterized by the expression of genes specifically up-regulated in this cell-type. We sought to construct a regulatory network that specifies ES cell-specific expression using binding sites of transcriptional regulators under the undifferentiated state. In order to infer regulatory interactions, we made use of published expression profiling data that compared undifferentiated with differentiating ES cells. The rationale is that nine (Nanog, Oct4, Sox2, Klf4, n-Myc, c-Myc, Esrrb, Zfx, Tcfcp2l1) out of the thirteen TFs we studied are known to be coordinately up-regulated in ES cells and their levels are down-regulated upon differentiation or in differentiated cell-types [62][63][64]. Two sets of published experiments were used to define genes that are differentially expressed during differentiation [62][65]. The use of two independently generated datasets minimizes biases in gene expression differences that are due to different ways of differentiating ES cells.

To reconstruct the transcriptional regulatory network, we need to determine the direct target genes of each transcription factor. The direct target genes must satisfy two criteria: a) genes up-regulated in mESC in relative to differentiated genes; b) the binding events of transcription factors in its proximal regions.

For criterion a), we computed fold-change values for each gene. We first applied a filter that removes probesets suffixed with “s_at” or “x_at”. If a gene still had more than one probeset associated with it, we used, in each comparison, the probeset that showed the greatest fold change. Genes were then ranked by the fold change values and those with >2 fold-change were identified to be up-regulated in mESC. The threshold of 2 fold-change is widely accepted in microarray gene expression analysis [62][65].

For criterion b), we started with the TF-gene association matrix proposed in Section 5.2, where each row corresponds to a gene, each column corresponds to a transcription factor, and the value at the i th row and the j th column represent the distance-dependent TF-gene association score of the i th gene and the j th TF. Suppose the TF-gene association matrix is M , we transformed M by multiplying it with the TF-TF covariance matrix, as follow:

$$M' = M\Sigma$$

where Σ is a 14×14 matrix that represent the pair-wise covariance among 13 TFs and the regulator Suz12. The transformed matrix M' incorporated the information of multiple TF co-localization. It is less sensitive to the false-positives in the prediction from a single ChIP-seq library, hence improved the reliability of gene-TF association.

Next, for each transcription factor, we ranked the genes in descending order of the gene-TF association represented in M' . The top-ranked genes are determined to satisfy criterion b). The thresholds for defining top-ranked genes in the two lists were determined empirically by maximizing the number of genes in the intersection subject to two constraints: there had to be at least twice as many genes in the intersection as the number expected by chance, and the null model (that there are no genes in excess) had to be rejected with $p < 10^{-3}$. This method allows us to make use of the unique features of our binding datasets (signal intensity and unbiased survey) and avoid the use of a single cutoff for all datasets.

The flowchart in Figure 5.7 summarizes the steps used to reconstruct the transcriptional regulatory networks. A network model based on the thirteen transcription factors and Suz12 as depicted in Figure 5.8 reveals both anticipated and novel aspects of the relationships between these transcription factors. Consistent with previous studies, this model shows regulatory feedback loops for Oct4, Sox2 and Nanog [48][59](Boyer et al., 2005; Chew et al., 2005; Loh et al., 2006). An interesting feature of this network is the inter-connectivity among eleven of the thirteen transcription factors being profiled.

Figure 5.7: The flowchart summarizing the steps used to reconstruct the transcriptional regulatory networks.

Figure 5.8: Network of regulatory interactions inferred from ChIP-seq dataset and from gene-expression changes during differentiation. Nodes are ChIP-seq-assayed transcription factors. Arrows point from the transcription factor to the target gene. Two sets of published experiments were used to define genes that are differentially expressed during differentiation [62][65]. Thick arrows represent interactions inferred from binding data and both expression experiments, whereas thin arrows represent interactions inferred from binding data and only one of the expression experiments.

Chapter 6

Discussion and Future Work

In this chapter, I first summarize the methods presented in the thesis work. Next, the limitations of the methods are discussed. Finally, I propose a few potential directions of future work.

6.1. Summary of Thesis Work

ChIP-seq is a powerful technique for the genome-wide study of protein-DNA interactions and histone modifications. Large amount of ChIP-seq datasets have been generated and published. The analysis of ChIP-seq data posed new challenges to the bioinformatics community. In the proposed thesis work, we developed three tools, Peak-finder, CCAT, and ChIPDiff, for identifying ChIP-enriched sites with different ChIP-seq experimental design. Based on the prediction of ChIP-enriched genomic sites, we reconstructed the key transcriptional regulatory network in mouse embryonic stem cells (mESC) from a large-scale ChIP-seq dataset.

Peak-finder was designed for the applications with a single ChIP library. It was derived from a peak-finding algorithm in ChIP-PET analysis. In this algorithm, the significance of a genomic location is measured by the number of DNA fragments overlapping that location. The difference between ChIP-PET and ChIP-seq is that ChIP-PET sequences

both ends of a DNA fragment while ChIP-seq sequences the 5' end only. Therefore, we proposed a concept of “virtual fragment” for ChIP-seq analysis, by which the genomic location of the 3' end of a fragment is approximated based on the fragment length distribution. A peak-calling approach was also developed to determine the cut-off threshold based on user-defined false discovery rate (FDR). In the peak-calling approach, we assumed a random model in which the DNA fragments are uniformly distributed in the whole genome. Peak-finder was successfully applied to ChIP-PET and ChIP-seq datasets. Wet-lab validation showed that the significance measure in Peak-finder is linearly correlated with the protein-DNA binding affinity (Figure. 2.4), and the threshold determined in the peak-calling procedure is reasonable (Table 2.2).

CCAT was designed for the applications with a ChIP library and a control library. The development of CCAT was initiated in our investigation of weak ChIP-seq signal. Several lines of evidence showed that there are intrinsic system biases in ChIP-seq technology. These biases introduce considerable false positives in the identification of weak ChIP-enriched sites with a single ChIP-seq library. Therefore, a control library needs to be included for the normalization against the intrinsic biases, and a new computational tool is required to facilitate the analysis. For this purpose, we proposed a linear signal-noise model for ChIP-seq analysis with negative control. The noise rate defined in the model provides a measurement of the data quality, which determines the capability of ChIP-seq to detect weak ChIP-enriched sites. We developed an iterative algorithm to estimate the noise rate using control library, and derived a library-swapping approach for FDR estimation. The performance of the proposed approaches was demonstrated with spike-in datasets. These approaches, integrated in the CCAT

framework, were further applied to H3K4me3 and H3K36me3 datasets, and identified significantly more weak sites than previous predictions. Two reasons counts for the improvement of sensitivity: firstly, the estimation of noise rate led to a better measurement of background noise level, which was overestimated by commonly used approach of “normalization against sequencing depth”; secondly, as shown in the spike-in simulation, the library-swapping approach improved the accuracy of FDR estimation, for which the previous p-value correction methods are too conservative. Our results on the histone modification datasets showed that the weak ChIP-seq signals may correspond to different genomic features from that of strong signals. Therefore, studying the weak signals could extend the scope of biological discoveries made from ChIP-seq data. Besides our example on histone modifications, another interesting topic is to study the weak transcription factor binding sites (TFBS), which possibly implicate the indirect binding mediated by TF-TF interactions. We expect the CCAT framework would facilitate such study in the future.

ChIPDiff was designed to compare two ChIP libraries that were generated from different cell types or treatments. Specifically, we emphasized on the identification of Differential Histone Modification Sites (DHMSs), which is of great interest in the elucidation of cell-specific regulatory mechanism in cancer research and stem cell research. Although the DHMSs could be identified using a naïve fold-change approach, the false positive rate is very high due to the variation of random sequencing. Therefore, we developed a Hidden Markov Model (HMM) based approach for the genome-wide quantitative comparisons of histone modifications. Using stem cell as a model system, we applied ChIPDiff to identify the DHMSs between ESC and NPC. We showed that ChIPDiff is able to render

robust and technically reproducible results though evaluation with K27 data. We further applied ChIPDiff to K4 and K36 data and achieved several interesting biological discoveries. The experimental results indicate that ChIPDiff is a useful tool for the comparative study of histone modifications between cell-types or different biological treatments.

The reconstruction of transcriptional regulatory networks is one of the most important application fields of ChIP-seq technology. Based on the prediction of ChIP-enriched genomic sites, we further developed an integrative analysis framework to reconstruct the transcriptional regulatory networks from large-scale ChIP-seq datasets and additional data sources such as microarray expression. We first employed *de novo* motif discovery tools and Monte Carlo simulation to determine the co-localization of multiple transcription factors. Next, we identify the putative target genes by calculating a TF-gene association matrix based on the genomic distance between the binding sites and the genes. In comparison to the conventional approaches using arbitrary threshold of distance, the TF-gene association matrix approach adapts to various patterns of TF concentration, and is more robust against the false positives in the binding site prediction. Finally, the list of putative target genes is refined based on gene expression datasets. In our experiment, we used the mESC cells as a model system, and applied the proposed framework to a large-scale ChIP-seq datasets that involves 13 TFs and 2 major regulators in mESC. As the result, we successfully reconstruct the core transcriptional regulatory networks of mESC, which is consistence to the previous biological knowledge.

6.2 Limitations

There are several limitations of the methods proposed in this thesis work.

● Peak-finder

Peak-finder is a tool for single library analysis. Therefore, one of the major drawbacks of Peak-finder is that the DNA fragments are assumed to be uniformly distributed in the random background model. Although the strong ChIP signals are still detectable with the uniform model, prediction of the weak sites is relatively difficult due to the intrinsic biases of ChIP-seq. In another tool CCAT, we proposed an alternative signal-noise model to solve this problem based a negative control library. Another limitation of Peak-finder lies on the representation of ChIP-enriched sites. In Peak-finder, the ChIP-enriched sites are defined in term of “peaks”, which refer to the summits in a ChIP-seq profile. For most applications on sequence-specific TF binding sites, the definition of “peaks” is reasonable, where the genomic locations of protein-DNA binding events can be represented in an accurate manner. However, for some applications with a broader ChIP-seq pattern such as the case in H3K36me3, “peaks” could not reflect the real pattern of ChIP enrichment. In these applications, ChIP-seq analysis needs to be performed by searching for the ChIP-enriched “regions” rather than “peaks”. Finally, the resolution of binding site prediction was not fully optimized. After the publication of Peak-finder, several tools such as SISSR [20] and MACs [24] significantly improved the resolution of binding site prediction.

• CCAT

A major drawback of CCAT is that only portion of the original dataset was utilized due to the re-sampling procedure. One solution to this problem is to use a bootstrapping strategy in the region scoring module. The option of bootstrapping is provided in the latest version of CCAT. In detail, the original dataset is re-sampled using different random seeds, and the scores of candidate regions are re-computed for a number of bootstrapping passes. The final score for a region is the median of the scores computed in all the bootstrapping passes. This can improve the sensitivity and specificity of the ranking list. A point of note is that the FDR could be slightly over-estimated when the bootstrapping strategy is used. In addition to the re-sampling CCAT requires the ChIP library and control library be prepared on the same cell-line, with the identical platform and configurations. This limited the application of CCAT to the comparison among different sources of datasets.

• ChIPDiff

Firstly, the bin size in ChIPDiff was set to be 1k bps in ChIPDiff, of which the resolution is relatively low when considering the nucleosome size of 200 bps (including the linker). The resolution, however, is limited due to the sequencing depth: if we reduce the bin size, there would not be enough fragment counts to be included in a bin for a reliable prediction. Secondly, the specificity, sensitivity and repeatability of our approach were evaluated based on technical replicates or a limited list of DHMSs inferred from biological knowledge and gene expression. There might be an argument on whether these data provide a “golden” standard for the evaluation. In fact, such a “golden” standard is very difficult to define for most biological data. Thirdly, we used the total number of

ChIP fragments for the normalization against sequencing depth. This normalization procedure is subject to the noise level of ChIP experiment. As an alternative, qPCR measurements [75] on a few “control” sites may provide a better way for normalization. And finally, in the pre-processing step, multiple tags retrieved from different fragments and mapped to the same genomic location were counted only once, which may result in error in quantitative measurement.

● **Reconstruction of transcriptional regulatory network**

In literature, the thesis work proposed in Chapter 5 is the first systematic computational method for the reconstruction of transcriptional regulatory networks based on ChIP-seq datasets. Therefore, there are several aspects pending improvement. Firstly, not all the binding site proximal to a gene is functional. In this sense, the identification of putative target genes can be improved by incorporating H3K4me1 histone modification sites, which is an enhancer mark and implies functional protein-DNA binding. Secondly, the determination of target genes based on distance is problematic. Recent studies showed the expression of a gene can be mediated by a remote regulatory element, which may be located more than 1M bps away from that gene, or even in a different chromosome [76][77][78]. New approaches need to be developed to determine the target genes not only based on genomic distance, but also based on higher-order chromatin structures. Thirdly, the protein-DNA binding energy is not quantitatively incorporated in the TF-gene association measure. In a recent study, Zhao *et al.* proposed a more complex quantitative model to describe the relationship between the binding energy and gene expression [79]. They applied the method on our mESC ChIP-seq datasets [14] and showed improvement of target gene prediction. Last but not the least, the transcriptional

regulatory networks can be extended by incorporating the epigenetic regulatory mechanisms, which plays an important role in stabilizing and “memorizing” the cell state.

6.3 Future Work

Due to the limitations mentioned in Section 6.2, the ChIP-seq analysis methods proposed in this thesis can be further improved. The future improvement will focus on the following five aspects. Firstly, “peak-finding” algorithms can be generalized into “region-finding” algorithms in Peak-finder and CCAT, since many biological problems are related to a span of ChIP-enriched genomic regions rather than the local maximum in ChIP-seq profile. It would be interesting to clearly define the boundary of ChIP-seq regions and to study the genomic elements associated to the boundaries. Secondly, the ChIP-seq signal is a mixture of several components rendered by multiple binding sites when these binding sites are close to each other (<200bps). To study the co-motifs related to multiple binding sites, de-convolution algorithms need to be developed to separate these components. Thirdly, ChIPDiff was designed to identify the differential histone modification sites. Is it possible to modify the HMM-based approach for the applications to differential binding sites? Fourthly, and the most importantly, the reconstruction of transcriptional regulatory networks is one of key problems in bioinformatics. An *ad hoc* approach to this problem was proposed in Chapter 5. The question is how to establish a computational framework towards a systematic solution using integrative analysis of ChIP-seq and gene expression? Based on our observation of the co-localization of multiple TFs, a possible direction is to re-define the regulators as “TF modules” rather

than a single TF. In this sense, deeper studies on the combinatorial effect of co-TFs and co-motifs are necessary. Finally, it is interesting to borrow the idea from the field of signal processing for the purpose of noise reduction in ChIP-seq analysis. Zhang *et al.* have successfully applied wavelet transform to determine the nucleosome positioning from ChIP-seq [82]. Although their work is limited to a specific application, it implies the potential of solving ChIP-seq related problems using conventional signal processing algorithms.

The genomic research is coming to an era of high-throughput sequencing (HTS). Besides ChIP-seq, quite a few HTS-based technologies have been developed to locate the genomic sites that are associated with various molecular features in the nucleus (genomic sequence, protein-DNA binding events, epigenetic markers, RNA expression, and splicing variants, *etc.*). However, an open question is how to determine the “functional sites”, which are crucial to elucidate the relationship between these molecular features and the phenotype. In the light of this, I will focus my future research on the development of integrative computational methods for the analysis of functional sites based on multiple sources of HTS data. By such study, I target at: a) to reveal the phenotype-associated transcriptional/post-transcriptional regulatory mechanism; b) to predict the phenotype from a limited set of nuclear molecular signatures. Although a specific topic is dependent on the research fund and data availability, there are three potential directions.

- **Comparative analysis of multiple HTS datasets**

In this study, I will employ signal processing techniques and statistical approaches for a comparative analysis among multiple HTS datasets that correspond to different

phenotypes. In the past, I developed ChIPDiff (Chapter 4) to identify the cell-specific histone modification sites based on pair-wise ChIP-Seq library comparison. This work is potentially extendable to deal with other types of HTS data (*e.g.*, whole-genome sequencing, RNA-Seq), different experimental design (*e.g.*, hierarchical sampling, time-course), as well as various molecular features (*e.g.*, copy number variation, protein-DNA binding, DNA methylation, alternative splicing).

- **Discovery of combinational modules of multiple molecular events**

The phenotype takes a combinatorial effect of multiple molecular events that occur concurrently or sequentially. These events can be observed and measured from HTS datasets in genome-wide scale. For instance, we discovered from ChIP-seq datasets that the co-localization of several key factors is crucial to determine the fate of mESC (Chapter 5). Therefore, a potential direction of my research is to define combinational modules of multiple molecular events using statistical and data-mining approaches. By this study, I hope I can make progress for the elucidation and modeling of the phenotype-associated transcription/post-transcriptional mechanism.

- **Predicting phenotype from molecular signatures**

This direction is of great potential for drug development, clinical diagnosis, and disease risk prediction. Previous study in this direction emphasize on the utilization of microarray [80][81]. These approaches only monitor the gene signature on the transcriptome level, but fail to consider other molecular events such as genome variation, protein-DNA binding, epigenetic modification, and RNA splicing. The advent of HTS techniques suggested more comprehensive approaches to determine the molecular signatures for

phenotype prediction. In my plan, there are three steps to achieve this goal. Firstly, the phenotype-associated molecular features will be identified in genome-wide scale from the HTS datasets; secondly, machine-learning approaches will be employed to assess the predictive power of the features; finally, an optimized set of features will be selected. By these steps, I target at an accurate and low-cost predictive technique for practical use.

References

- [1] Kim T.H. and Ren B. (2006) “Genome-wide analysis of protein-DNA interactions.” *Annu. Rev. Genom. Hum. Genet.*, **7**:81-102.
- [2] Lieb J.D. *et al.* (2001) “Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association”. *Nature Genetics*, **28**:327-334.
- [3] Ren B. *et al.* (2000) “Genome-wide location and function of DNA binding proteins.” *Science*, **290**:2306-2309.
- [4] Iyer V.R. *et al.* (2001) “Genomic binding sites of the yeast cell-cycle transcription facts SBF and MBF.” *Nature*, **409**:533-538.
- [5] Kim T.H. *et al.* (2005) “A high-resolution map of active promoters in the human genome.” *Nature*, **436**:876–880.
- [6] Roh T. *et al.* (2004) “High-resolution genome-wide mapping of histone modifications”. *Nature Biotech.*, **22**:1013-1016.
- [7] Wei C. L. *et al.* (2006) “A global map of p53 transcription-factor binding sites in the human genome”. *Cell*, **124**:207-219.
- [8] Wheeler D.A. *et al.* (2008) “The complete genome of an individual by massively parallel DNA sequencing.” *Nature*, **452**:872-876.
- [9] Nagalakshmi U. *et al.* (2008) “The transcriptional landscape of the yeast genome defined by RNA sequencing.” *Science*, **320**:1344-1349.

- [10] Cokus S.J. *et al.* (2008) "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." *Nature*, **452**:215-219.
- [11] Lieberman-Aiden E. *et al.* (2009) "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science*, **326**:289-293.
- [12] Barski A. *et al.* (2007) "High-resolution profiling of histone methylations in the human genome." *Cell*, **129**: 823-837.
- [13] Johnson D.S. *et al.* (2007) "Genome-wide mapping of in vivo protein-DNA interactions." *Science*, **316**:1497-1502.
- [14] Chen X. *et al.* (2008) "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell*, **133**:1106-1117.
- [15] Marson A. *et al.* (2008) "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." *Cell*, **134**:521-533.
- [16] Loh Y.H., *et al.* (2006) "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." *Nature Genetics*, 38:431-440
- [17] Boyle A.P. *et al.* (2008) "F-Seq: a feature density estimator for high-throughput sequence tags." *Bioinformatics*, **24**:2537-2538.
- [18] Robertson G. *et al.* (2007) "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." *Nature Methods*, **4**:651-657.

- [19] Valouev A. *et al.* (2008) “Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.” *Nature Methods*, **5**:829-834.
- [20] Jothi R. *et al.* (2008) “Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.” *Nucleic Acids Res.*, **36**:5221-5231.
- [21] Mikkelsen T.S. *et al.* (2007) “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.” *Nature*, **448**:553-560.
- [22] Zang C. *et al.* (2009) “A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.” *Bioinformatics*, **25**:1952-1958.
- [23] Ji H. *et al.* (2008) “An integrated software system for analyzing ChIP-chip and ChIP-seq data.” *Nature Biotech.*, **26**:1293-1300.
- [24] Zhang Y. *et al.* (2008) “Model-based Analysis of ChIP-seq (MACS).” *Genome Biology*, **9**:R137.
- [25] Vega V.B. *et al.* (2009) “Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries.” *PLoS One*, **4**:e5241.
- [26] Rozowsky J. *et al.* (2009) “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.” *Nature Biotech.*, **27**:66-75.
- [27] Tuteja G. *et al.* (2009) “Extracting transcription factor targets from ChIP-Seq data.” *Nucleic Acids Res.*, **37**:e113.

- [28] Xu H. *et al.* (2010) “A signal-noise model for significance analysis of ChIP-seq with negative control.” *Bioinformatics*, **26**:1199-1204.
- [29] Xu H. *et al.* (2008) “Genome-wide Identification of Differential Histone Modification Sites from ChIP-seq Data.” *Bioinformatics*, **24**:2344-2349.
- [30] Zhao X. D. *et al.* (2007) “ Whole genome mapping of histone H3 Lys 4 and 27 trimethylation reveals distinct genomic compartments in human embryonic stem cells”. *Cell Stem Cell*, **1**:286-298.
- [31] Hogg R.V. and Craig A.T. (1978) *Introduction to Mathematical Statistics*, 4th edition. New York: Macmillan.
- [32] Choi S.C and Wette R. (1969) “Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias.” *Technometrics*, **11**:683–690
- [33] Qi Y. *et al.* (2006) “High-resolution computational models of binding events”. *Nature Biotech.*, **23**:963-970.
- [34] Metropolis N. and Ulam S. (1949) “The Monte Carlo Method.” *J. Amer. Statistical Assoc.*, **44**: 335–341.
- [35] Santos-Rosa H. *et al.* (2002) “Active genes are tri-methylated at K4 of histone H3.” *Nature*, **419**:407-411.

- [36] Ng H.H. *et al.* (2003) "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." *Mol. Cell*, **11**:709-719.
- [37] Schneider R. *et al.* (2004) "Histone H3 lysine 4 methylation patterns in higher eukaryotic genes." *Nat. Cell Biol.*, **6**:73-77.
- [38] Felsenfeld G. *et al.* (1996) "Chromatin Structure and Gene Expression." *Proc. Natl. Acad. Sci. USA*, **93**:9384-9388.
- [39] Cook E.H. and Scherer S.W. (2008) "Copy-number variations associated with neuropsychiatric conditions". *Nature*, **455**: 919–23.
- [40] Meissner A. *et al.* (2008) "Genome-scale DNA methylation maps of pluripotent and differentiated cells." *Nature*, **454**:766-770.
- [41] Nix D.A. *et al.* (2008) "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks." *BMC Bioinformatics*, **9**:523.
- [42] Storey J. D. (2003) "The positive false discovery rate: a Bayesian interpretation and the q-value." *Ann. Statist.*, **31**:2013-2035.
- [43] Marson A. *et al.* (2008) "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." *Cell*, **134**:521-533.
- [44] Benjamini Y. & Hochberg Y. (1995) "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *J. R. Stat. Soc. Ser. B.*, **57**:289-300.

- [45] Guenther M.G. *et al.* (2007) "A chromatin landmark and transcription initiation at most promoters in human cells." *Cell*, **130**:77-88.
- [46] Heintzman N.D. *et al.* (2007) "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." *Nature Genet.* **39**:311–318.
- [47] Martin C. and Zhang Y. (2005) "The diverse functions of histone lysine methylation." *Nature Rev. Mol. Cell Biol.*, **6**:838-849.
- [48] Boyer L.A. *et al.* (2006) "Polycomb complexes repress developmental regulators in murine embryonic stem cells." *Nature*, **441**:349-353.
- [49] Bernstein B.E. (2006) "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." *Cell*, **125**:315-326.
- [50] Widschwendter M. *et al.* (2007) "Epigenetic stem cell signature in cancer." *Nature Genet.*, **39**:157-158.
- [51] McGarvey K.M. *et al.* (2006) "Silenced tumor suppressor genes reactivated by DNA demethylation do not return to a fully euchromatic chromatin state." *Cancer Res.* **66**:3541-3549.
- [52] Quackenbush J. (2002) "Microarray data normalization and transformation". *Nature Genet.*, **32**:496-501.

- [53] Rabiner L. (1989) “A tutorial on hidden Markov models and selected applications in speech recognition.” *Proc. IEEE*, **77**:257-286.
- [54] Gan Q. *et al.* (2007) “Concise review: epigenetic mechanism contribute to pluripotency and cell lineage determination of embryonic stem cells.” *Stem Cell*, **25**:2-9.
- [55] Gupta A.K. and Nadarajah S. (ed.) (2004) *Handbook of beta distribution and its applications*. Marcel Dekker, New York.
- [56] Raiffa H. and Schaifer R. (1961) *Applied Statistical Decision Theory*. Harvard University Press.
- [57] Li W. *et al.* (2005) “A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences.” *Bioinformatics*, **21 Suppl 1**:i274-i282.
- [58] Welch L.R. (2003) “Hidden Markov Models and the Baum-Welch Algorithm.” *IEEE Information Theory Society Newsletter*, **53**:1-1.
- [59] Loh Y.H., *et al.* (2006) “The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.” *Nature Genetics*, **38**:431-440
- [60] Pavese G. *et al.* (2001). “An algorithm for finding signals of unknown length in DNA sequences.” *Bioinformatics*, **17 Suppl 1**:S207–S214.
- [61] Eisen M. B. *et al.* (1998) “Cluster analysis and display of genome-wide expression patterns.” *Proc. Natl. Acad. Sci. USA*. **95**:14863-14868.

- [62] Ivanova, N.B., *et al.* (2002). “A stem cell molecular signature.” *Science*. **298**:601–604.
- [63] Ivanova, N. B. *et al.* (2006). “Dissecting self-renewal in stem cells with RNA interference.” *Nature*. **442**:533–538.
- [64] Ramalho-Santos M. *et al.* (2002). “Stemness: transcriptional profiling of embryonic and adult stem cells.” *Science*. **298**:597–600.
- [65] Zhou, Q. *et al.* (2007) “A gene regulatory network in mouse embryonic stem cells.” *Proc. Natl. Acad. Sci. USA*. **104**:16438–16443.
- [66] Chew J.L. *et al.* (2005). “Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells.” *Mol. Cell. Biol.* **25**:6031–6046.
- [67] Akutsu T. *et al.* (1999) “Identification of genetic networks from a small number of gene expression patterns under the Boolean network model.” *Pacific Symp. Biocomp.* **99**: 17-28.
- [68] Hoon M. *et al.* (2003) “Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pacific Symp. Biocomp.* **8**:17-28.
- [69] Friedman N. *et al.* (2000) “Using Bayesian network to analyze expression data” *J. Comp. Biol.* **7**:601-620.

- [70] Imoto S. *et al.* (2002) “Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression.” *Pacific Symp. Biocomp.* **7**:437-449.
- [71] Imoto S. *et al.* (2002) “Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network.” *Proc. 1st IEEE Computer Society Bioinformatics Conference*, 210-227.
- [72] Hu Z. *et al.* (2007) “Genetic reconstruction of a functional transcriptional regulatory network”. *Nature Genetics*, **39**:683-687.
- [73] Pilpel Y. *et al.* (2001) “Identifying regulatory networks by combinatorial analysis of promoter elements.” *Nature Genetics*, **29**:153-159.
- [74] Segal E. *et al.* (2003) “Module networks: Identifying regulatory modules and their condition specific regulators from gene expression data.” *Nature Genetics*, **34**:166-176.
- [75] Ding C. and Cantor C.R. (2004) “Quantitative analysis of nucleic acids – the last few years of progress.” *J. of Biochem. And Mol. Bio.*, **37**:1-10.
- [76] Kurukuti S. *et al.* (2006) “CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2.” *Proc. Natl. Acad. Sci. USA*, **103**:10684-10689.
- [77] Splinter E. *et al.* (2006) “CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus”. *Genes Dev.*, **20**:2349-2354.

- [78] Hou C. *et al.* (2008) “CTCF-dependent enhancer-blocking by alternative chromatin loop formation.” *Proc. Natl. Acad. Sci. USA*, **105**:20398-20403.
- [79] Zhao Y. *et al.* (2009) “Inferring binding energies from selected binding sites” *PLoS Comput. Biol.*, **5**:e1000590
- [80] Ohali A. *et al.* (2005) “Prediction of high risk Ewing’s sarcoma by gene expression profiling.” *Oncogene*, **23**:8997-9006.
- [81] Kim B.S. *et al.* (2005) “Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer.” *Bioinformatics*, **21**:517-528.
- [82] Zhang Y. *et al.* (2008) “Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq.” *BMC Genomics*, **9**:537.

APPENDIX I: Gene Ontology Associated with H3K4me3 and H3K27me3 Sites in Section 2.5

Table S1: Ontology for Genes with H3K4me3 and no H3K27me3

Over represented	p-value	Under represented	p-value
<u>Nucleoside, nucleotide and nucleic acid metabolism</u>	1.70E-13	<u>Olfaction</u>	3.63E-49
<u>Intracellular protein traffic</u>	1.81E-10	<u>Chemosensory perception</u>	6.73E-46
<u>Protein metabolism and modification</u>	3.43E-09	<u>Cell surface receptor mediated signal transduction</u>	1.18E-43
<u>Protein biosynthesis</u>	3.28E-08	<u>Sensory perception</u>	6.42E-43
<u>DNA metabolism</u>	2.37E-07	<u>Signal transduction</u>	4.61E-41
<u>Cell cycle</u>	2.78E-06	<u>G-protein mediated signaling</u>	5.03E-38
<u>General vesicle transport</u>	3.21E-05	<u>Cell communication</u>	2.18E-30
<u>Pre-mRNA processing</u>	9.34E-05	<u>Immunity and defense</u>	1.26E-24
<u>DNA repair</u>	2.30E-04	<u>Ligand-mediated signaling</u>	8.81E-18
<u>mRNA splicing</u>	3.76E-04	<u>Developmental processes</u>	5.78E-15
<u>Protein modification</u>	7.36E-04	<u>Cytokine/chemokine mediated immunity</u>	8.68E-13
<u>DNA replication</u>	1.19E-02	<u>Steroid hormone metabolism</u>	1.18E-11
<u>Mitosis</u>	1.33E-02	<u>Cytokine and chemokine mediated signaling pathway</u>	2.43E-11
<u>Chromatin packaging and remodeling</u>	3.78E-02	<u>Ectoderm development</u>	7.42E-11
<u>Protein targeting and localization</u>	5.32E-02	<u>Cell adhesion</u>	9.27E-10
<u>Protein folding</u>	7.66E-02	<u>Macrophage-mediated immunity</u>	1.27E-09
		<u>Cell adhesion-mediated signaling</u>	2.44E-09
		<u>Ion transport</u>	2.63E-09
		<u>Natural killer cell mediated immunity</u>	3.35E-09

Table S2: Ontology for Genes with both H3K4me3 and H3K27me3

Over represented	p-value	Under represented	p-value
<u>Ectoderm development</u>	2.94E-79	<u>DNA metabolism</u>	1.14E-07
<u>Neurogenesis</u>	2.60E-56	<u>Chemosensory perception</u>	5.11E-07
<u>mRNA transcription regulation</u>	1.89E-55	<u>Olfaction</u>	1.50E-06
<u>mRNA transcription</u>	9.94E-48	<u>Protein biosynthesis</u>	1.17E-03
<u>Neuronal activities</u>	2.36E-40	<u>DNA repair</u>	2.77E-03
<u>Signal transduction</u>	2.39E-33	<u>Intracellular protein traffic</u>	2.52E-02
<u>Cell communication</u>	7.24E-31	<u>Proteolysis</u>	2.98E-02
<u>Mesoderm development</u>	1.72E-25	<u>DNA replication</u>	6.36E-02
<u>Segment specification</u>	2.39E-19	<u>Nuclear transport</u>	7.70E-02
<u>Skeletal development</u>	2.67E-18	<u>Mitosis</u>	9.29E-02
<u>Nucleoside, nucleotide and nucleic acid metabolism</u>	2.06E-14	<u>Pre-mRNA processing</u>	9.41E-02
<u>Synaptic transmission</u>	2.26E-14		
<u>Other neuronal activity</u>	1.05E-13		
<u>Cell surface receptor mediated signal transduction</u>	1.11E-13		
<u>Ion transport</u>	3.21E-11		
<u>Cation transport</u>	1.49E-10		
<u>Cell proliferation and differentiation</u>	2.23E-10		
<u>Receptor protein tyrosine kinase signaling pathway</u>	8.25E-09		
<u>Anterior/posterior patterning</u>	1.67E-08		
<u>Ligand-mediated signaling</u>	6.34E-08		
<u>Embryogenesis</u>	1.04E-07		
<u>Other receptor mediated signaling pathway</u>	7.11E-07		
<u>Transport</u>	4.51E-06		
<u>Oncogenesis</u>	1.13E-05		

Table S3: Ontology for Genes with no H3K4me3 and no H3K27me3

Over represented	p-value	Under represented	p-value
<u>Olfaction</u>	5.85E-43	<u>Nucleoside, nucleotide and nucleic acid metabolism</u>	3.63E-72
<u>Chemosensory perception</u>	8.44E-43	<u>mRNA transcription</u>	2.98E-37
<u>G-protein mediated signaling</u>	1.50E-38	<u>mRNA transcription regulation</u>	3.47E-32
<u>Sensory perception</u>	5.36E-37	<u>Intracellular protein traffic</u>	1.46E-17
<u>Cell surface receptor mediated signal transduction</u>	1.21E-35	<u>Protein biosynthesis</u>	9.53E-17
<u>Immunity and defense</u>	1.14E-34	<u>Cell cycle</u>	2.25E-15
<u>Signal transduction</u>	1.42E-25	<u>Protein modification</u>	1.17E-14
<u>Cytokine and chemokine mediated signaling pathway</u>	6.87E-15	<u>Protein metabolism and modification</u>	3.70E-12
<u>Macrophage-mediated immunity</u>	1.39E-12	<u>General vesicle transport</u>	2.05E-11
<u>Cell communication</u>	2.25E-12	<u>Cell cycle control</u>	5.82E-11
<u>Cytokine/chemokine mediated immunity</u>	4.26E-12	<u>DNA metabolism</u>	1.83E-10
<u>Steroid hormone metabolism</u>	4.15E-11	<u>mRNA splicing</u>	6.53E-09
<u>Natural killer cell mediated immunity</u>	5.70E-10	<u>Protein phosphorylation</u>	7.72E-08
<u>Ligand-mediated signaling</u>	9.71E-10	<u>Pre-mRNA processing</u>	7.79E-08
<u>Cell adhesion</u>	4.32E-08	<u>DNA repair</u>	2.28E-07
<u>Steroid metabolism</u>	7.50E-08	<u>Neurogenesis</u>	2.65E-06
<u>T-cell mediated immunity</u>	8.13E-07	<u>DNA replication</u>	7.41E-05
<u>Cell adhesion-mediated signaling</u>	1.11E-06	<u>Cell proliferation and differentiation</u>	1.21E-04
<u>B-cell- and antibody-mediated immunity</u>	1.71E-05	<u>Other intracellular signaling cascade</u>	8.26E-04
<u>Interferon-mediated immunity</u>	4.54E-05	<u>rRNA metabolism</u>	2.69E-03
<u>Proteolysis</u>	4.15E-04		
<u>Complement-mediated immunity</u>	8.98E-04		
<u>Blood circulation and gas exchange</u>	2.10E-03		
<u>Cell structure</u>	3.26E-03		
<u>Transport</u>	3.59E-03		

APPENDIX II: Proof of Proposition 1 in Section 3.2

Proposition 1: Given a genomic region $r = [a, b]$, and $r \subset \bar{E}$, the observation vectors $\mathbf{c}_r = [c_a^+, c_a^-, \dots, c_b^+, c_b^-]^T$ and $\mathbf{d}_r = [d_a^+, d_a^-, \dots, d_b^+, d_b^-]^T$ follow the same multivariate distribution under the hyperprior π , if $M = \alpha N$.

Proof:

Note that the following assumptions have been made in the signal-noise model:

- (a) $c_a^+, c_a^-, \dots, c_b^+, c_b^-, d_a^+, d_a^-, \dots, d_b^+, d_b^-$ are independent observations;
- (b) $\forall \theta, \forall x, c_x^\theta \sim \text{Poisson}(u_x^\theta M)$;
- (c) $\forall \theta, \forall x, d_x^\theta \sim \text{Poisson}(v_x^\theta N)$;
- (d) $\forall \theta, \forall x, v_x^\theta = (1 - \alpha)s_x^\theta + \alpha n_x^\theta$;
- (e) $[u_1^+, u_1^-, \dots, u_L^+, u_L^-]^T$ and the noise vector $[n_1^+, n_1^-, \dots, n_L^+, n_L^-]^T$ follow the same multivariate hyper distribution π ;
- (f) $\forall \theta, \forall x \in \bar{E}, s_x^\theta = 0$.
- (g) $\mathbf{r} \subset \bar{E}$;
- (h) $M = \alpha N$.

We denote $\mathbf{u}_r = [u_a^+, u_a^-, \dots, u_b^+, u_b^-]^T$, $\mathbf{v}_r = [v_a^+, v_a^-, \dots, v_b^+, v_b^-]^T$, based on (a), (b), and (c), we have:

$$\forall \mathbf{y} \in \mathcal{R}^{2b-2a+2}, \forall \mathbf{z} \in \mathcal{R}^{2b-2a+2}$$

$$P(\mathbf{c}_r = \mathbf{y} | M\mathbf{u}_r = \mathbf{z}) = P(\mathbf{d}_r = \mathbf{y} | N\mathbf{v}_r = \mathbf{z}) \quad (\text{S1})$$

We also denote $\mathbf{n}_r = [n_a^+, n_a^-, \dots, n_b^+, n_b^-]^T$, based on (e), the distributions of $\mathbf{u}_r$ and $\mathbf{n}_r$ under π are identical, which is equivalent to:

$$\forall \mathbf{z} \in \mathcal{R}^{2b-2a+2}$$

$$P(M\mathbf{u}_r = \mathbf{z}|\pi) = P(M\mathbf{n}_r = \mathbf{z}|\pi)$$

(S2)

From (d), (f), (g), and (h), we have:

$$N\mathbf{v}_r = \alpha N\mathbf{n}_r = M\mathbf{n}_r$$

(S3)

Considering Eq S1, S2, and S3, we have:

$$\forall \mathbf{y} \in \mathcal{R}^{2b-2a+2}$$

$$\begin{aligned} P(\mathbf{c}_r = \mathbf{y}|\pi) &= \int_{\mathbf{z}} P(\mathbf{c}_r = \mathbf{y}|M\mathbf{u}_r = \mathbf{z})P(M\mathbf{u}_r = \mathbf{z}|\pi)d\mathbf{z} \\ &= \int_{\mathbf{z}} P(\mathbf{d}_r = \mathbf{y}|N\mathbf{v}_r = \mathbf{z})P(M\mathbf{n}_r = \mathbf{z}|\pi)d\mathbf{z} \\ &= \int_{\mathbf{z}} P(\mathbf{d}_r = \mathbf{y}|N\mathbf{v}_r = \mathbf{z})P(N\mathbf{v}_r = \mathbf{z}|\pi)d\mathbf{z} \\ &= P(\mathbf{d}_r = \mathbf{y}|\pi) \end{aligned}$$

(S4)

E.O.P

APPENDIX III: Proof of, and Notes on Proposition 2 in Section 3.4

Proposition 2: Under the conditions:

b) $\Pr(f(\mathbf{d}_r, \mathbf{c}_r) > t \mid r \notin \bar{E}) < \Pr(f(\mathbf{c}_r, \mathbf{d}_r) > t \mid r \notin \bar{E})$, and

b) $M = \alpha N$;

The pFDR for a threshold t can be approximated by:

$$pFDR(t) = \frac{\#\{f(\mathbf{d}_r, \mathbf{c}_r) > t\}}{\#\{f(\mathbf{c}_r, \mathbf{d}_r) > t\}}$$

where the operator $\#\{\Delta\}$ is defined to be the total number of regions satisfying the condition Δ .

Proof:

Rewriting the pFDR representation in Eq. 3.7 in Section 3.4:

$$pFDR(t) = \Pr(r \in \bar{E} \mid f(\mathbf{c}_r, \mathbf{d}_r) > t) \quad (S5)$$

and considering Proposition 1 in Appendix II and condition (b) in this proposition, we derived:

$$\begin{aligned} pFDR(t) &= P(\mathbf{r} \in \bar{E} \mid f(\mathbf{c}_r, \mathbf{d}_r) > t) \\ &= \frac{P(f(\mathbf{c}_r, \mathbf{d}_r) > t \mid \mathbf{r} \in \bar{E})P(\mathbf{r} \in \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \\ &= \frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t \mid \mathbf{r} \in \bar{E})P(\mathbf{r} \in \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \end{aligned} \quad (S6)$$

Based on condition (a), we have:

$$\frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t | \mathbf{r} \notin \bar{E})P(\mathbf{r} \notin \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \ll \frac{P(f(\mathbf{c}_r, \mathbf{d}_r) > t | \mathbf{r} \notin \bar{E})P(\mathbf{r} \notin \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \leq 1 \quad (\text{S7})$$

Therefore:

$$\begin{aligned} pFDR(t) &\approx \frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t | \mathbf{r} \in \bar{E})P(\mathbf{r} \in \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \\ &+ \frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t | \mathbf{r} \notin \bar{E})P(\mathbf{r} \notin \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \\ &= \frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t)}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)} \end{aligned} \quad (\text{S8})$$

which can be approximated by counting the frequency:

$$pFdr(t) \approx \frac{\#\{f(\mathbf{d}_r, \mathbf{c}_r) > t\}}{\#\{f(\mathbf{c}_r, \mathbf{d}_r) > t\}} \quad (\text{S9})$$

E.O.P.

In Eq. S8, we made the approximation since the absolute value of the compartment $\frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t | \mathbf{r} \notin \bar{E})P(\mathbf{r} \notin \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)}$ is small. By this, we controlled the *absolute error* of estimation. The absolute error here is defined as the magnitude of the difference between the exact value and the approximation, i.e.:

$$e_{abs} = |\hat{x} - x|$$

where x is the exact value, and $\hat{x}$ is an approximation of x .

However, a small absolute error does not guarantee a small *relative error*, which is defined as the ratio of the absolute error to the magnitude of the exact value:

$$e_{rel} = \frac{|\hat{x} - x|}{|x|}$$

This problem happens especially when the compartment $\frac{P(f(\mathbf{d}_r, \mathbf{c}_r) > t | \mathbf{r} \in \bar{E}) P(\mathbf{r} \in \bar{E})}{P(f(\mathbf{c}_r, \mathbf{d}_r) > t)}$ is small, hence it invoked a question: should we control the absolute error or the relative error when estimating empirical FDR?

First of all, we agree that the relative error is more preferred in many applications. However, for empirical FDR estimation of ChIP-seq, the relative error is difficult to control. Note that the empirical FDR is computed as the ratio of the frequency of called regions under empirical random model to the frequency of observations, where the frequency is subject to the variation of Bernoulli process. In the Proof of Proposition 2, the step of “frequency counting” is represented by Eq. S9. To control the relative error to a low level, a large value of the numerator in the ratio, i.e. the frequency by random, is required. Unfortunately, this is not practical in real ChIP-seq applications since we don’t want to include too many false positives.

Here we give an example to further clarify this issue. In an experiment, 10,000 sites are called given a threshold t , and the error introduced in Eq. S8 is zero, we want to know the range of absolute error and relative error introduced in Eq. S9. For this purpose, we compute the upper-bounds (3-stdev) of errors, conditional on $Fdr(t)$, as show in Table S4.

In practice, the exact value of $Fdr(t)$ is unknown. Based on the table, we are still confident that the absolute error in Eq. S9 is under control with upper bound $< 2\%$. As

long as we are able to minimize the absolute error in Eq. S8, we can guarantee that the overall absolute error is also controlled. Meanwhile, the relative error in Eq. S9 could be large when $Fdr(t)$ is small. For such a case, minimizing the relative error in Eq. S8 would be useless since majority of the relative error could be introduced in Eq. S9. Therefore, controlling the absolute error, rather than controlling the relative error, is more practically sound in the empirical FDR estimation of ChIP-seq.

Table S4: Comparison of the upper-bound (3 standard deviation) of absolute error and relative error in empirical FDR estimation.

$Fdr(t)$	3*stdev, absolute error	3*stdev, relative error
0.001	0.000978	0.977789
0.002	0.001386	0.693108
0.005	0.002163	0.432543
0.01	0.003019	0.301918
0.02	0.004291	0.214531
0.05	0.006397	0.127945
0.1	0.008937	0.089372
0.2	0.011786	0.058929
0.5	0.014788	0.029575

APPENDIX IV: Quantitative PCR Experiment for H3K4me3 site validation in Section 3.7.

To validate reliability of ChIP-Seq sites identified using our methods, we performed chromatin immunoprecipitation coupled with quantitative real time PCR. Chromatin extract was prepared from cross-linked murine E14 embryonic stem cells and used to perform chromatin immunoprecipitation using antibody against H3K4m3 (Abcam, ab8580, [21]) as described in [14]. 15 positive sites were randomly selected from the peaks in the H3K4me3 regions exclusively predicted by CCAT; 10 negative sites were selected from the genomic regions outside the predicted regions. DNA sequences were extracted from 250 bps to the left and right of the peaks. Primers were then designed from these peaks using online software PrimerFinder at <http://qpcr.probefinder.com/roche3.html>. Quantitative real time PCR was done in duplicate. See Table S5 and S6 for the details of qPCR information.

Table S5: 15 Positive Sites for qPCR validation

peak region	left primer	right primer	qPCR enrichment fold	standard deviation
chr10:84886275-84886775	aactcaacctccgctctcg	aggcagttgcagcacaaaag	9.200536	1.43214483
chr4:137150475-137150975	aggtccctgtacctgtgg	cataacggagcctcaccttg	7.244847	1.38971197
chr14:68125075-68125575	cacaagcagtccttttcg	ggaaagagcgttacttacgtg	10.88447	2.23468925
chr1:186543775-186544275	ccgggaggaccttaaacact	cccatatttgagggtttgg	5.93748	0.51811613
chr2:118223775-118224275	gactgcattcctggtgatcc	tcgccctgaacagacaaga	18.37292	2.52602059
chr3:16683225-16683725	gcctgtgttcttactggttt	aaggcaattgcttaaggaaaaa	39.66654	3.46138018
chr3:8921025-8921525	gctggagagccaatccaa	ccaagcaggctttctctcag	4.556017	0.23039136
chr4:135511175-135511675	ggacctcctctccgtcct	tctcacctgcccatcagg	7.415493	0.71497678
chr17:53048825-53049325	ggctgtcgggtaagagaggt	ccatggcactaaacagaagga	15.46403	4.41098887
chr2:172488675-172489175	tagatcccgcatactctgc	ctcgtgggtctgagttaccac	25.78454	5.17800413
chr11:94062325-94062825	tatggggcctcagcagat	agatgctgggtgggtcacg	7.991132	4.01628753
chr2:59641675-59642175	tctatctcaatctgtcgcgatg	actgccaagaactgggaca	1.279997	0.19343427
chr11:98584275-98584775	tctgtggggaagacgacag	ctccgtgtcctccgtgtc	28.75579	6.16175149
chr7:29040525-29041025	tttgggctctaggagaaaaaca	gacagagccaggactcaagc	7.764835	1.48945641
chr7:6019475-6019975	tttggtgcatgtgtgcatt	cctgttgaggataccacac	29.99624	3.30351778

Table S6: 10 Negative Sites for qPCR validation

primer region	left primer	right primer	qPCR enrichment fold	standard deviation
chr1:25439725-25439925	ccctaactcacagcaccaca	atgtgtccttcacccagat	1.107647	0.38322063
chr10:29174511-29174673	agcttattctcaaaaagaccaga	gaccagcagctaattggttg	0.833239	0.01149672
chr11:5759207-5759415	ccaggattggcctctcataa	aaggctccttctctttgcag	2.16248	0.42454176
chr12:44958375-44958575	tcactgagaaaattggtgaagtaa a	acacattaagtctgcctttga a	1.024943	0.18736502
chr13:68266025-68266225	gaagagaaatgttgaattcagg	gccaaaccataattacatcac a	1.074212	0.24459159
chr15:86549775-86549975	agcaagtaataactcccttggac	agccctgagtcaggta	0.89494	0.15955988
chr15:5592214-5592428	gatggtggcttaccaggtagaa	aatccaagagcagcatgagg	0.748822	0.2037187
chr16:18868569-18868769	agccttctataccatgagctt	gagagctttggaggacca	0.596254	0.07926273
chr16:34929075-34929275	gtagccctctcctgctgat	catccaggagctaggtttg	1.062134	0.4436199
chr16:91068775-91068975	ggctgggtcaaaagaatgag	ccttcaggcagttgagaat	0.975071	0.57205069