

Biological Insights of Transcription Factor through Analyzing ChIP-Seq Data

by

Kaida Ning

B.Sc., Zhejiang University, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

The Faculty of Graduate Studies
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

November 2009

© Kaida Ning 2009

Abstract

ChIP-Seq is a technology for detecting in vivo transcription factor binding sites or histone modification sites on a genome wide scale. How to utilize the large scale data and find out biological insights is a challenging question for us.

Here, we analyzed three ChIP-Seq data sets for human HeLa cell, including data of a transcription factor called STAT1, data of RNA polymerase II (Pol2), and data of histone monomethylation (Me1). With these data sets, we looked into the spacial relationship between STAT1 binding sites, Pol2 binding sites, Me1 flanked regions and the gene transcription start sites; we checked the intersection of locations of STAT1 binding sites, Pol2 binding sites and Me1 flanked regions; we did de novo motif discovery for the sequences around the STAT1 binding sites, and predicted several transcription factors whose binding sites may form cis-regulatory module with STAT1 binding site; we put the STAT1-centered sequences into different categories based on their spacial relationship with Pol2 binding sites and Me1 flanked regions, and found that the de novo discovered motifs' occurrence rates are different in sequences of different categories; we also analyzed the ChIP-Seq data along with gene expression data, and found that STAT1 binding may be related with genes' differential expression under IFN-gamma stimulation.

We suggest that further ChIP-Seq experiment be carried out for TFs corresponding to the de novo predicted motifs, and that gene expression be characterized for the IFN-gamma stimulated HeLa cell on the whole genome scale.

Table of Contents

Abstract	ii
Table of Contents	iii
List of Tables	vii
List of Figures	viii
Abbreviation of Terminology	x
Acknowledgements	xi
Dedication	xii
1 Introduction	1
1.1 Transcription factors	1
1.1.1 Gene expression in a nutshell	1
1.1.2 Role of transcription factor (TF) in regulating gene expression	1
1.1.3 DNA motif for transcription factor	2
1.1.4 De novo motif discovery	2
1.1.5 Database for motifs of known TFs	4
1.2 ChIP-Seq experiment	4
1.2.1 The procedure of ChIP-Seq experiment	4
1.2.2 ChIP-Seq's advantage over ChIP-chip	5
1.2.3 Identifying TF binding site through analyzing ChIP-Seq data	5
1.2.4 Current methods of comparing the accuracy of software for analyzing ChIP-Seq data	9
1.3 ChIP-Seq data of STAT1, Pol2 and Me1	9
1.3.1 STAT1	10
1.3.2 RNA Polymerase II	10

Table of Contents

1.3.3	Me1	11
1.4	GO analysis of genes potentially regulated by a pair of TFs .	12
1.4.1	The GO project	12
1.4.2	Finding out the enriched GO terms (biological process) for genes potentially regulated by a pair of TFs	12
1.5	Overall goals and significance	13
1.5.1	What we expect to know by analyzing the spacial relationship between TSS, STAT1, Pol2 and Me1 . . .	13
1.5.2	Predicting the TFs that may collaborate with STAT1 in regulating the gene transcription	14
1.5.3	Relating microarray data with ChIP-Seq data	14
2	Methods	15
2.1	ChIP-Seq data set	15
2.1.1	STAT1, Pol2 ChIP-Seq data	15
2.1.2	Predicted Me1 binding sites from ChIP-Seq data . . .	15
2.2	De novo motif discovery for sequences around STAT1 binding sites	15
2.2.1	Obtaining DNA sequences around high confidence STAT1 binding sites	15
2.2.2	Handling the repeating regions	16
2.2.3	De novo motif discovery with GADEM	16
2.2.4	Visualization of de novo predicted motifs	17
2.3	Obtaining gene information data set	19
2.4	Relating STAT1 binding with DE genes	20
2.4.1	Genes differentially expressed in IFN-gamma stimulated HeLa cell	20
2.4.2	Genes differentially expressed in other three types of IFN-gamma stimulated human cells	20
2.4.3	Checking STAT1 binding site in promoter region of DE genes	21
2.5	Multiple test for proportions	22
2.5.1	Testing proportion of two samples	22
2.5.2	Multiple test with Bonferroni adjustment	22
3	Results	23
3.1	Analyzing ChIP-Seq data	23
3.1.1	Prediction of STAT1, Pol2 binding sites from ChIP-Seq data	23
3.1.2	Prediction of Me1 flanked regions of biological interest	23

Table of Contents

3.2	Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS	24
3.2.1	Relationship between the length of region upstream or downstream of TSS and the number of regions having STAT1, Pol2 binding site and Me1 flanked region . .	26
3.2.2	The inconsistency in the increase in number of regions having STAT1, Pol2 binding sites and Me1 flanked regions with respect to the increase in the length of regions	30
3.2.3	Intersection of locations of STAT1 binding sites, Pol2 binding sites and Me1 flanked regions	35
3.3	De novo motif discovery for STAT1 sequences	38
3.3.1	Frequency of the binding sites corresponding to de novo discovered motifs occurring in the STAT1 sequences	38
3.3.2	Visualizing the de novo discovered motifs	40
3.3.3	Location overlapping issue of binding sites corresponding to de novo detected motifs	45
3.4	Selection of motif _x whose corresponding TFs are most likely to cooperate with STAT1	47
3.5	Literature review for the cooperation of STAT1 and selected motif _x	52
3.5.1	Cooperation of STAT1 and Nanog	52
3.5.2	Cooperation of STAT1 and HEB	53
3.5.3	Cooperation of STAT1 and Tel2 (or TFs from the ETS family)	54
3.5.4	Cooperation of STAT1 and AP1	54
3.6	Genome-wide analysis on function of genes potentially regulated by STAT1 and TF _x	56
3.7	The occurrence rate of binding sites corresponding to de novo discovered motifs	62
3.7.1	Categories of STAT1 sequences	62
3.7.2	Occurrence rate of binding sites corresponding to de novo motifs in STAT1 sequence of different categories	62
3.7.3	Occurrence rate of binding sites corresponding to de novo motifs in STAT1 sequence of different categories (categories based on Me1 flanked regions and Pol2 sites with more stringent criteria)	65
3.8	The occurrence rate of binding sites corresponding to different combinations of de novo discovered motifs	65

Table of Contents

3.9	Relating STAT1 binding sites with DE genes	68
3.9.1	DE genes detected on Chromosome 22 of IFN-gamma stimulated HeLa cell	68
3.9.2	DE genes detected for other three types of human cells under IFN-gamma stimulation	68
3.9.3	Intersection of DE genes in HeLa cell and DE genes in other three types of cells	69
3.9.4	Proportion of DE gene promoters having STAT1 bind- ing sites and proportion of all gene promoters having STAT1 binding sites	69
4	Discussion	73
4.1	The binding sites of TFs	73
4.1.1	The overlap of binding site locations of de novo pre- dicted motifs	73
4.1.2	What are the non-coding regions in the genome? Are TF binding in regions far from genes regulating the gene transcription?	73
4.2	Uncertainty in the specificity of TF binding	74
4.2.1	Uncertainty in deciding which TF has binding site similar to the de novo predicted motif	74
4.2.2	Uncertainty in the specificity of how TFs collaborate	74
4.3	The condition of TF binding	75
4.4	Potential wet-lab experiment	75
4.4.1	Potential ChIP-Seq experiment	75
4.4.2	Potential gene expression experiment	76
4.5	Use of R, Perl and SQL	76
	Bibliography	78
 Appendices		
A	Parameter setting for running MACS	84
B	Parameter setting for running DecoyMasker	85
C	Parameter setting for running GADEM	86

List of Tables

1.1	Summary of ChIP-Seq data analysis software	8
3.1	Summary of ChIP-Seq reads and binding site predicted for STAT1 and Pol2	24
3.2	De novo motifs predicted for 401bp sequences around the STAT1 binding sites.	41
3.3	Five selected de novo motifs predicted for 401bp sequences around the STAT1 binding sites.	49
3.4	High confidence HEB regions with STAT1 GAS binding site on Chromosome 19	53
3.5	Total number of genes in each group of genes potentially reg- ulated by STAT1 and TF _x	57
3.6	Comparison of each motif's occurrence rate in STAT1 se- quences of different categories.	64
3.7	Summary of number of total genes and number of DE genes in three microarray data sets	69
3.8	Proportion of promoter regions with STAT1 binding sites . .	71
3.9	Genes differentially expressed in all 3 types of cells and whether there is STAT1 binding site in their promoter regions	72

List of Figures

1.1	Main steps of ChIP-Seq experiment: chromatin immunoprecipitation and sequencing.	6
1.2	The short reads from ChIP-Seq mapped back to genome. . .	6
2.1	WebLogo representation of STAT1	19
3.1	Distribution of the distance between nearby Me1 binding sites	25
3.2	Relationship between the length of upstream or downstream region of TSS and the number of regions with at least one STAT1 binding site.	27
3.3	Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Pol2 binding site.	28
3.4	Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Me1 binding site.	29
3.5	Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Pol2 binding site detected without control data.	31
3.6	Increase of number of upstream or downstream region of TSS having STAT1 with respect to the region length.	32
3.7	Increase of number of upstream or downstream region of TSS having Pol2 with respect to the region length.	33
3.8	Increase of number of upstream or downstream region of TSS having Me1 flanked region with respect to the region length. .	34
3.9	Intersection of STAT1 binding sites, all Pol2 binding sites and all Me1 flanked regions.	36
3.10	Intersection of STAT1 binding sites, top 10K Pol2 binding sites and top 20K Me1 flanked regions.	37
3.11	Number of times the binding sites of de novo motifs are found in 9,992 401-bp STAT1 sequences.	39

List of Figures

3.12	Number of times the location of two motifs overlapping with each other.	46
3.13	The overlapping of the STAT1 binding site and the NF- κ B binding site in the genomic region of -5.8kb in iNOS promoter	47
3.14	An example of three motifs whose predicted binding sites are overlapping with each other	48
3.15	The color and corresponding p-value of enriched term used by GOrilla.	58
3.16	Biological process significantly enriched in genes which are potentially regulated by STAT1 and Nanog	58
3.17	Biological process significantly enriched in genes which are potentially regulated by STAT1 and HEB	59
3.18	Biological process significantly enriched in genes which are potentially regulated by STAT1 and Tel2	60
3.19	Biological process significantly enriched in genes which are potentially regulated by STAT1 and AP1	61
3.20	Biological process significantly enriched in genes which are potentially regulated by STAT1	61
3.21	Motif occurrence rate in STAT1 sequences of different categories (categories are based on all Pol2 binding sites and all Me1 binding sites).	63
3.22	Motif occurrence rate in STAT1 sequences of different categories (categories are based on top Pol2 binding sites and top Me1 binding sites).	66
3.23	Occurrence rate of binding site corresponding to different motif combinations in STAT1 sequences belonging to different categories	67
3.24	Intersection of DE genes found in three types of cells after IFN-gamma stimulation.	70

Abbreviation of Terminology

DAG: directed acyclic graph

DE gene: differentially expressed gene

GO: Gene Ontology

iNOS: inducible nitric oxide synthase gene

Me1: monomethylation of Lys4 of histone H3

Pol2: RNA polymerase II

PWM: position weight matrix

STAT1: signal transducers and activators of transcription 1

TF: transcription factor

TSS: transcription start site

Acknowledgements

I am very grateful to Dr. Raphael Gottardo and Dr. Gordon Robertson, who supervised me and gave me many good suggestions on this thesis work.

Thanks to those who have helped me or accompanied me during the two-year study. Thousands of miles away from hometown, I can not imagine how to make all these happen without you.

Thanks to CIHR/MSFHR Bioinformatics Training Program. I believe, being a student in this program will be a treasured experience for the rest of my life.

Dedication

For my parents.

Chapter 1

Introduction

1.1 Transcription factors

Transcription factors (TFs) are proteins which regulate the transcription of a gene by binding to its promoter or enhancer region (Blackwood et al. [7], Latchman [36]). There are about 25,000 genes in the human cell, and more than 2,000 of these genes encode TFs (Babu et al. [2], Messina et al. [45], Pennisi [50]).

1.1.1 Gene expression in a nutshell

In human haploid genome, there are totally just over 3 billion DNA base pairs. The information of genes is encoded within DNA.

Gene expression is the process by which a gene guides the production of functional gene product. It includes two steps: the genetic information being transferred to mRNA via transcription, and then, mRNA guides the synthesis of protein via translation.

1.1.2 Role of transcription factor (TF) in regulating gene expression

A TF has specific 3-D DNA binding domain, which lets it recognize and bind to specific DNA sequence. And TFs are usually classified on the basis of their DNA binding domains, such as Zinc finger domain, Homebox domain, Ets domain, etc (Latchman [36], Pabo [48]).

While many transcription factors can bind to diverse genic and intergenic genomic locations, when they bind to upstream of the transcription start site (TSS) of a gene they can be associated directly with changes in the genes expression levels (i.e., the transcription level of mRNA). TFs usually cooperate with each other collectively in regulating the gene expression in eukaryotic cell (Berman et al. [6], Davidson [13]).

Within a TF, there are domains effecting activation or repression on recruitment of basal transcriptional complex, so that the TF can promote

(as an activator) or block (as a repressor) a gene's transcription. The activation domains function by either stimulating the assembly of the basal transcriptional complex or stimulating its activity once it has assembled. The basal transcriptional complex is composed of RNA polymerase II and various transcription factors, such as TFIIB and TFIID (Roeder [55]). The repression domains function either by interfering with the action of a positively acting factor (indirect repression) or by interacting directly with the basal transcriptional complex (direct repression) (Latchman D.S. [37]).

1.1.3 DNA motif for transcription factor

DNA binding sites are distinctive short DNA sequences which can be recognized and bound by specific TFs. The pattern of a set of recurring short DNA sequences of one binding site is called "motif" (D'haeseleer [49]).

There is variability in sites for a motif, and a motif for a specific TF is usually well conserved in most of the positions, but not in all the positions. For example, it is known that in yeast, although the consensus binding site for the TATA binding protein (TBP) is TATAAAA, a wide variety of A/T-rich sequences, such as TATATAT or TATATAA, can function as TATA boxes and can interact with TBP (Chen et al. [11], Singer et al. [63]). Moreover, different binding sites for a TF have different affinity in binding with the TF (Bulyk et al. [10]).

For that reason, it is not appropriate to simply represent a motif with a fixed DNA sequence. Instead, a motif is usually represented with position weight matrix (PWM), where each column represent a position in the motif, and each entry of the matrix is the occurrence rate of A, C, G and T at a specific position. A motif can also be visualized in a way which is easy for human to recognize a motif's component and conservation level at each position (Crooks et al [12], Gorodkin et al. [22]). For example, WebLogo is a popular software for motif visualization (Crooks et al [12]). We will further discuss the PWM and WebLogo representation in the method section "Visualization of de novo predicted motif".

1.1.4 De novo motif discovery

Summary of de novo motif discovery from sequences

The basic idea of de novo motif discovery is to identify one or more sets of well conserved, and over-represented subsequences, from a set of DNA sequences, and at the same time identify the motif corresponding to these

subsequences. “De novo” means, the motif discovery is done with no prior knowledge of the composition of the motif.

De novo motif discovery is often carried out as an important step for studying gene regulation: for instance, it can discover over-represented motifs that are common to promoters of genes with similar expression patterns. After discovering the motifs, we can tell which TFs have binding sites like the de novo discovered motifs, and predict these TFs as regulators of the genes under study. Our prediction can be used as “prior knowledge” for wet lab experiments carried out to identify the true TFs that regulate the genes.

Overview of software available for de novo motif discovery

Many software are freely available for de novo motif discovery, for example, AlineACE (Roth et al. [56]), Gibbs Motif Sampler (Thompson et al. [69]) and MEME (Bailey et al. [3, 4]) are some of the well known software, and GADEM (Li [39]) is a recently developed software we used for this thesis work. EM (expectation maximization) and Gibbs sampling are the two major algorithms used by de novo motif discovery tools, the former learns the latent variable and PWM through EM updating (applied by GADEM, MEME, etc.), the latter learns PWM through Gibbs sampling (applied by AlineACE, Gibbs Motif Sampler, etc.).

Details about GADEM in de novo motif prediction

GADEM can be viewed as an extension of the well known motif discovery tool MEME. We chose to use it because it is fast, user friendly and gives competitive prediction for motifs. The main step GADEM takes in predicting motif is summarized in the following:

- GADEM runs in the unit of cycle and each cycle contains several generations. Within each cycle, GADEM does the following things:
 - finds out top-ranked k-mers from the sequences;
 - initializes the PWMs using spaced dyad, which is two over-represented k-mers separated by spacers;
 - uses EM iteration to update the PWMs and latent variable, stop iteration when the likelihood converges or the number of iteration reaches user defined number;

- after EM iteration, transforms the PWM so that each entry of it is integer, and use the integerized PWM to scan for binding sites in the sequences; and declare a subsequence as binding site when the p-value of its score with respect to the PWM is below a threshold;
- aligns the binding sites declared to have the same motif; calculates entropy score and corresponding statistical significance for the alignment, and use it as the fitness score for the spaced dyad from which the PWM is initialized;
- mutate or crossover all the spaced dyads except for the best performing ones, and train PWM based on them again;
- repeats the previous steps until the maximal number of generations is reached;
- outputs motifs and masks them in the original data.

1.1.5 Database for motifs of known TFs

There are several databases containing motifs of known TFs, such as TRANSFAC (Wingender et al. [71]) database, JASPAR (Sandelin et al. [58]) database, et al. And, in order to compare the similarity between de novo discovered motifs and the motifs of known TFs, we can use tool such as STAMP (Mahony [43]), which provides an interface for uploading PWM of de novo motif and searching for the motif in the database which is most similar to the de novo motif.

1.2 ChIP-Seq experiment

A ChIP-Seq experiment is initially designed to identify the *in vivo* binding sites of a TF on the whole genome scale, which is a key step in understanding gene regulation. ChIP-Seq is also used to identify RNA Polymerase binding sites and histone modification sites on the whole genome scale.

1.2.1 The procedure of ChIP-Seq experiment

Here, we summarize how ChIP-Seq is used to find out binding sites of a TF.

ChIP-Seq is an combination of chromatin immunoprecipitation (ChIP) and next-generation sequencing. In the ChIP part of the process, TFs are cross-linked to the DNA in the cell, then the DNA is sheared into small

fragments by sonication. After that, an antibody which binds specifically to the TF to be studied will be added to the solution of DNA fragments, so that the DNA fragments bound by the TF to be studied are precipitated. In the sequencing part of the process, DNA fragments and the TFs are reverse-corss linked, and DNA fragments are sent to next generation sequencing machine. Short reads at one or both end(s) of DNA fragments are sequenced and the reads are then mapped onto genomic locations by a read-alignment algorithm (Mardis [44]). A ChIP-Seq experiment typically generates tens of millions of short reads during each instrument run.

In the ChIP-Seq experiment, the DNA fragments bound by transcription factor are the most frequently sequenced fragments, and short reads obtained for them form peaks at TF binding sites when mapped back to the genome. Given the short reads being mapped back to genome, TF binding sites are predicted at regions where the short reads are enriched.

We also know that ChIP-Seq characterizes the DNA fragments in the immunoprecipitated 'reagent', and this will contain not just protein-bound DNA fragments but also other ('background' or 'noise') fragments.

Figure 1.1 illustrates the main steps of ChIP-Seq and Figure 1.2 shows a schematic diagram of short reads mapped back to the genome, at a region with two TF binding sites.

1.2.2 ChIP-Seq's advantage over ChIP-chip

The main difference between ChIP-Seq and ChIP-chip is that the former sequences the protein-bound DNA fragments, the latter hybridizes the protein-bound DNA fragments to a tiling microarray.

The main advantages of ChIP-Seq over ChIP-chip includes higher spacial resolution and less input material requirement as well as independence from design and manufacture of tiled microarray (Mardis [44], Robertson et al. [53]).

1.2.3 Identifying TF binding site through analyzing ChIP-Seq data

Summary of identifying TFBS through analyzing ChIP-Seq data

Through analyzing ChIP-Seq data, we can detect genomic regions that are enriched for immuno-precipitated DNA fragments, and predict them as TF binding sites. We some times call enrichment profiles 'peaks' because the accumulated short reads look like peaks.

1.2. ChIP-Seq experiment

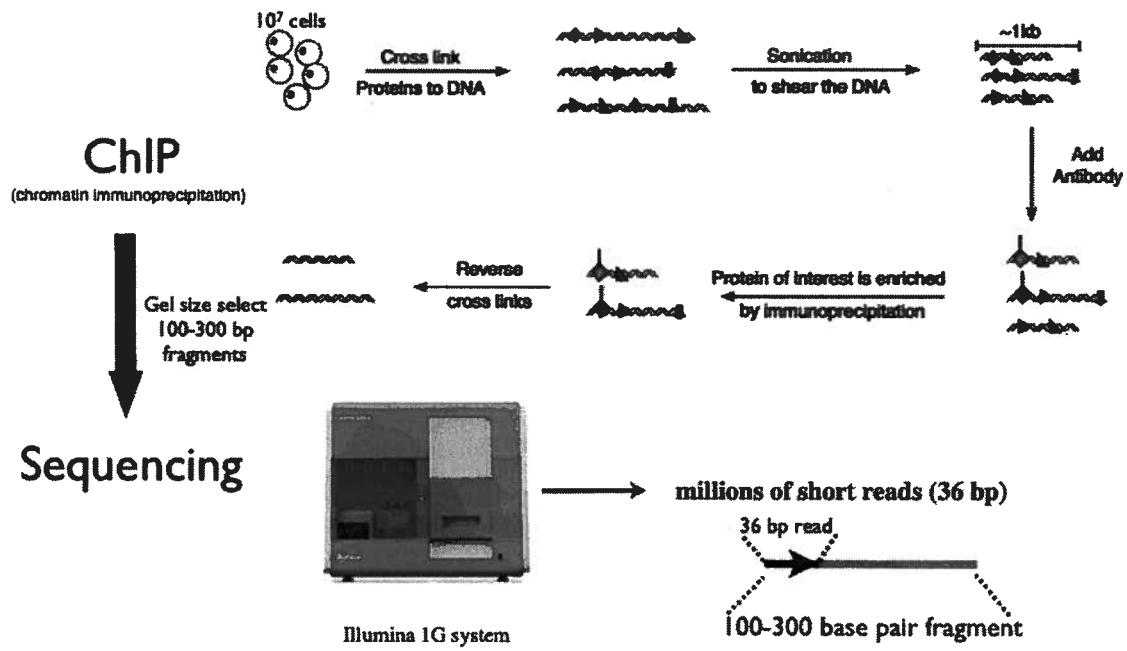


Figure 1.1: Main steps of ChIP-Seq experiment: chromatin immunoprecipitation and sequencing.

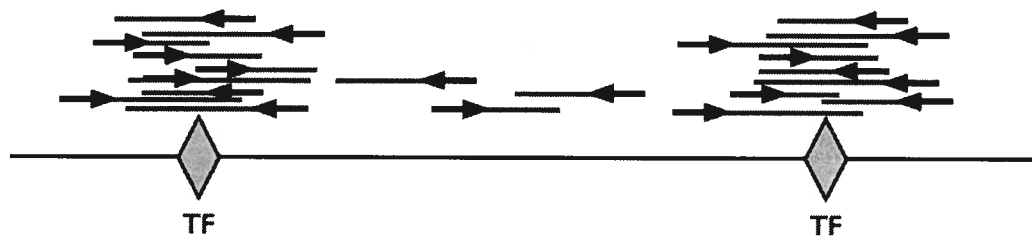


Figure 1.2: The short reads from ChIP-Seq mapped back to genome.

The common way to detect binding sites in ChIP-Seq data is to first scan the genome with a sliding window and select regions having more read counts than a user defined cutoff. The binding sites will then be predicted for those regions.

Distinguishing and removing false positive enriched regions is also necessary. ChIP-Seq data usually comes with control data. With control data, the signal in the ChIP-Seq sample and the signal in corresponding region of a control sample can be compared. A binding site will be called if the signal coming from ChIP-Seq is significantly stronger than that coming from control sample at the same region, and a local p-value will be calculated. Most of the peak-finding software authors claim that performance will be better when a control sample is analyzed with the ChIP-Seq sample, and if there is no control sample to use, background will be simulated.

Overview of software for analyzing ChIP-Seq data

There are several software available for analysing the ChIP-Seq data and predicting TF binding sites, such as CisGenome (Ji et al. [27]), FindPeaks (Fejes et al. [16]), MACS (Zhang et al. [74]), PICS (Zhang et al. [73]) and QuEST (Valouev et al. [70]). All are freely available.

The major difference between these software is in the way they model the ChIP-Seq reads: CisGenome model the merged forward and reverse reads with either a Poisson distribution or a negative binomial distribution (they suggest that the negative binomial model performs better); FindPeaks does not assume any distribution for the reads; MACS uses Poisson distribution to model the merged forward and reverse reads; PICS uses t-distributions to jointly model the forward and reverse reads, where each pair of close-by peaks are linked with a parameter telling the distance between them; QuEST is based on kernel density estimates of the forward and reverse reads separately.

Since a short read can come from either end of a sonicated DNA fragment, the short reads around a true binding site show a bimodal enrichment pattern. The peak-finding software merge the two peaks formed by forward/reverse short reads into one peak in either arbitrary or more statistical way, and the binding site is predicted as the summit of merged peak.

Besides the difference in modeling short reads, the peak-finders are also different in shifting the reads toward the real binding site and detecting peak summits, using control data and calculating FDR. We summarize and compare the features of them all in Table 1.1.

1.2. ChIP-Seq experiment

Table 1.1: Summary of ChIP-Seq data analysis software

Software	Density profile for short reads	Peak shift method	Use of control data	Global FDR control
CisGenome	negative binomial distribution or Poisson distribution	shift size is half of the medium distance between paired peaks	use all the control data as background if available	direct FDR control
FindPeaks	no specific density profile	Not available	use all the control data as background if available	no direct FDR control
MACS	Poisson distribution	shift size is half of the average distance between the high quality forward and reverse peaks	use all the control data as background if available	direct FDR control
PICS	t-distribution	shift size for each pair of peaks is based on the density model of the paired peaks	use all the control data as background if available	direct FDR control
QuEST	kernel density estimate	shift size is half of the average distance between forward peaks and reverse peaks	half of the control used as background, half of the control used as pseudo ChIP-Seq data	direct FDR control

Details about MACS in analyzing ChIP-Seq data

In our analysis, we chose to use MACS, because it is fast and easy to use. The main steps of MACS in analyzing ChIP-Seq data is summarized in the following:

- slides a window across the genome to identify regions where number of reads from ChIP-Seq sample is *mfold* more than number of reads from the control sample;
- randomly selects top 1,000 high quality peaks, and calculates the average distance between the forward peaks and reverse peaks (*d*);
- shifts all reads $d/2$ toward 3';

1.3. ChIP-Seq data of STAT1, Pol2 and Me1

- models the reads in the whole control data, and get a uniform parameter of Poisson distribution for control data, λ_{BG} ;
- for each ChIP-Seq region selected by sliding window, calculates λ_{local} based on λ_{BG} and the distribution of reads in the corresponding control sample, calculate the p-value of the region with respect to λ_{local} ;
- call a binding site in the region if p-value of the region is less than default cutoff (i.e., 10^{-5});
- for a region with called binding site, extend every read position d bases from its center, and use the location with the highest fragment pileup as the precise binding site;
- uses the same p-value cutoff on ChIP-Seq data and control data, and calculates FDR for the regions as Number of control peaks / Number of ChIP peaks.

1.2.4 Current methods of comparing the accuracy of software for analyzing ChIP-Seq data

On one hand, software for analyzing ChIP-Seq data is still new, and is continuing to evolve. On the other hand, little is known about the true TF binding sites, and the research community lacks a convincing way of evaluating the performance of peak detection software.

One empirical method for comparing different peak detection software is the false discovery rate (FDR), which is calculated as number of control peaks / number of ChIPSeq peaks detected at the same threshold.

Another method for comparison is checking the motif occurrence within certain distance of the peak center detected, the higher occurrence implies the better performance of the peak detection software (Zhang et al. [74]).

1.3 ChIP-Seq data of STAT1, Pol2 and Me1

ChIP-Seq can be used to identify RNA Polymerase binding sites, histone modification sites as well as TF binding sites on the whole genome scale. In the thesis, we analyzed the ChIP-Seq data for STAT1 (a transcription factor), RNA polymerase II and histone monomethylation.

1.3.1 STAT1

Signal transducer and activator of transcription 1 (STAT1) is a well studied transcription factor which is involved in IFN-dependent and growth factor-dependent signaling (Ramana [52]). One of the pathways STAT1 participate in is IFN- γ - Janus tyrosine kinase (JAK) - STAT1: Under normal conditions, STAT1 binds to the IFN- γ receptor, which locates on the cytoplasm. When cell gets stimulated by IFN- γ , the conformation of IFN- γ receptor changes, which causes the phosphorylation of JAK1 and JAK2, and later the phosphorylation of STAT1 (Schroder et al. [59]).

Most of the time, the phosphorylated STAT1 form homodimer and translocate from cytoplasm into the nucleus. The homodimer activates or represses transcription primarily by binding to IFN-gamma activation site (GAS) elements and it regulates the gene expression through collaboration with other TFs. Also, STAT1 can form heterodimer (STAT1 with STAT2) and bind to interferon-stimulated response elements (ISREs) with IRF-9 (Schroder et al. [59]).

A ChIP-Seq experiment indicated that in the IFN- γ stimulated cell, the STAT1 bound sites are about fourfold more than in the unstimulated cell (Robertson et al. [53]).

1.3.2 RNA Polymerase II

RNA Polymerase II (Pol2) is a protein of 515k Daltons. It is able to bind to the TATA region of the gene, unwind DNA, synthesize RNA according to the DNA when sliding through the DNA, and rewind DNA (Brooker [9]). Pol2 alone is not capable of recognizing the TATA region. To recognize the promoter and initialize transcription, Pol2 works with general transcription factors, such as TFIIB and TFIID, and together, they form the basal transcription complex (Kornberg [33]).

Pol2 not only accumulates at the actively transcribed gene region, but also accumulates at inactive genes. For example, β -globin gene is actively transcribed in immature cell, while the transcription can not be detected in the mature cell. Gariglio et al. found that Pol2 are evenly distributed along the β -globin gene in immature cell, while Pol2 accumulates at the promoter region of β -globin gene in the mature cell (Gariglio [21]).

It was reported that Pol2 can also bind to distant control elements, such as enhancers, and eventually associate with co-activators and general transcription factors. For example, using an *in vitro* reconstructed nucleosomal PSA enhancer, Louie et al. showed that Pol2 can be recruited to the en-

hancer independent of the promoter (Louie et al. [40]). In a paper by Blackwood et al., the mechanism that Pol2 or TFs recruited at a distant enhancer acting on the promoter by DNA looping or tracking along the chromatin was discussed (Blackwood et al. [7]).

Pol2 also binds to intergenic regions. For example, a ChIP-chip experiment on stationary phase (SP) yeast and mid-log (ML) yeast showed that Pol2 in SP is more predominantly located on inter gene regions (IGRs), whereas Pol2 in ML is more predominantly located on gene coding regions (GCRs); also, Pol2 is found at some GCRs in SP yeast, which facilitates the rapid transcriptional engagement when the SP exits (Radonjic et al. [51]).

1.3.3 Me1

Nucleosomes are the fundamental structures of chromatin. The nucleosome core particle is approximately 146 base pairs of DNA wrapping around a histone octamer, which is composed of two copies of the core histones: H2A, H2B, H3 and H4. The histone octamer and DNA is stabilized by the linker histone H1 (Luger et al. [42]).

Eukaryotic gene transcription is accompanied by acetylation and methylation of nucleosomes near promoters and enhancers. For example, Heintzman et al. carried out a ChIP-chip experiment on ENCODE regions, revealing that the TSS of active promoters are marked by both monomethylation and trimethylation of Lys4 of histone H3 (Me1, Me3), whereas enhancers are marked by Me1, not Me3 (Heintzman et al. [25]); a recent analysis of Me1, Me3, TF STAT1 in HeLa cell, and TF FOXA2 in mouse liver cell showed that Me1, not Me3, is the dominant modification for STAT1 sites and FOXA2 sites far away from the TSS (Robertson et al. [54]). Result from these two experiments also indicated that there were bimodal distribution of histone modification centered around the Pol2 binding sites or the TF binding sites, e.g., for the regions centered around the Pol2 binding sites or the TF binding sites, the modification signals are always detected about 200-1,000bp away from each other. In our analysis, we focused on the Me1 flanked regions, which are the regions between two Me1 sites located about 200-1,000bp from each other.

1.4 GO analysis of genes potentially regulated by a pair of TFs

1.4.1 The GO project

The Gene Ontology (GO) project provides an ontology of defined terms representing gene product properties. Three structured vocabularies (ontologies) have been developed in a species-independent manner, and they describe gene products in terms of their associated biological processes, cellular components and molecular functions separately (Ashburner et al. [1]).

The structure of GO is directed acyclic graph (DAG). In GO, each annotation is a node in the DAG, and an annotation may have more than one parent and have more than one child. The more we know about a gene product, the deeper its annotation lies in the DAG.

1.4.2 Finding out the enriched GO terms (biological process) for genes potentially regulated by a pair of TFs

By checking whether there is binding site for a specific TF in the promoter region of a gene, we can predict whether the gene is potentially regulated by that TF.

We know that in eukaryotic cell, the regulation of gene is typically achieved by several TFs which bind onto its promoter region simultaneously. Therefore, a gene is potentially regulated by one or more TFs if there is binding site(s) for the TF(s) in its promoter region.

For a group of genes which are potentially regulated by one or more TFs, we can find out whether there are GO terms significantly enriched in these genes, i.e., whether the genes potentially regulated by the TF(s) code proteins participating in the same biological processes more often than randomly selected genes. The finding of enriched GO terms can give support that the genes studied are really regulated by the same TF(s).

Several software packages are available to do this, such as GOrilla (Eden et al. [15]) and Ontologizer (Bauer et al. [5]).

The basic idea for finding out enriched GO term in n target genes is: 1) given n genes, within which b genes have a specific GO term; 2) given a set of N genes as background, which has B genes with that specific GO term; 3) assume that the n target genes are randomly sampled from these N genes, and calculate the probability that at least b out of n genes have that specific GO term.

1.5. Overall goals and significance

Using hypergeometric distribution for solving the problem is a popular solution (Bluthgen et al. [8]), which is illustrated in the formula below:

$$p(x \geq b) = \sum_{x=b}^{\min(n,B)} \frac{\binom{B}{x} \binom{N-B}{n-x}}{\binom{N}{n}} \quad (1.1)$$

For a specific GO term, if $p(x \geq b)$ is smaller than the given significance level in the target genes, it is declared to be enriched in the target genes.

1.5 Overall goals and significance

1.5.1 What we expect to know by analyzing the spacial relationship between TSS, STAT1, Pol2 and Me1

TF binding, Pol2 binding and Me1 flanked regions are all related to gene transcription: the binding of a TF to the promoter region can regulate the corresponding gene's transcription by either stimulating / repressing the assembly of the basal transcriptional complex or stimulate / repress its activity once it has assembled; Pol2 is a must for the protein coding gene transcription, it slides along the DNA when synthesizing RNA from DNA; the Me1 flanked region is found in the promoter and enhancer of actively transcribed genes (Heintzman et al. [25], Robertson et al. [54]).

The transcription start site (TSS) is the site within a gene where the transcription of DNA into RNA begins. Immediately upstream to TSS, there is gene promoter region, where the transcription factors and basal transcriptional complex bind to (Kutach et al. [35], Ohler et al. [47]).

As far as we know, there is no paper talking about how the TF binding sites, Pol2 binding sites and Me1 flanked regions are located with respect to the TSS (where the transcription begins) in the human genome. Here we checked, whether they all tend to occur in the upstream of TSS instead of in the downstream of TSS, and whether they are all located very close to the TSS.

We are also interested in how the locations of TF binding sites, Pol2 binding sites and Me1 flanked regions are related with each other in the genome. Through studying the ChIP-Seq data sets for STAT1, Pol2 and Me1 together, we can check the proportion of locations of STAT1 binding sites, Pol2 binding sites and Me1 flanked regions intersecting with each other, i.e., to which extent, these three factors are related with each other. Moreover, we will put STAT1 binding sites into different categories according to their

spacial relationship with Pol2 and Me1 flanked regions, and we will check whether the motif occurrence rate is different in different categories.

1.5.2 Predicting the TFs that may collaborate with STAT1 in regulating the gene transcription

We know that in eukaryotes, the TFs always collaborate with each other in regulating the gene transcription. Here, we want to find out which TFs may collaborate with STAT1 in regulating the gene transcription through de novo motif discovery for sequences around STAT1 binding sites. We hope that our prediction will give a direction to the potential ChIP-Seq experiment for other TFs in the future, so that the ChIP-Seq experiment is more oriented and money can be saved for the experiment of more interest.

We will verify our prediction through the literature review and the GO analysis of the genes potentially regulated by STAT1 and the predicted TFs.

1.5.3 Relating microarray data with ChIP-Seq data

Under normal conditions, the amount of various proteins in the cell is in an equilibrium, and together, the proteins make the cell function normally. When a cell gets stimulated (e.g. chemical treatment, temperature change, exposure to X-ray), the equilibrium inside it no longer exists: the mRNA level of certain genes will change, which leads to the change in amount of corresponding proteins. Microarray analysis is a technique for monitoring mRNA level on a genomic scale.

We know that the differential expression is caused by the binding of transcription factor in promoter region. Therefore, we want to check whether the STAT1 binding is related with genes' differential expression after IFN-gamma stimulation, i.e., whether it occurs more often in the promoters of genes whose expression level change after IFN-gamma stimulation than in the promoters of all the genes.

Chapter 2

Methods

2.1 ChIP-Seq data set

2.1.1 STAT1, Pol2 ChIP-Seq data

The STAT1 and Pol2 ChIP-Seq data with controls were published by Rozowsky et al.: STAT1 ChIP-Seq was done for the whole genome of IFN-gamma stimulated HeLa cell, while Pol2 ChIP-Seq was done for whole genome of unstimulated HeLa cell (Rozowsky et al. [57]). The data set was downloaded from:

<http://www.gersteinlab.org/proj/PeakSeq/>

2.1.2 Predicted Me1 binding sites from ChIP-Seq data

The Me1 data set we used was generated by Robertson et al., who carried out a genom-wide ChIP-Seq for Me1 in IFN-gamma stimulated HeLa cell (Robertson et al. [54]). Their Me1 ChIP-Seq data did not come with control. MACS method claims to have better performance when control data is provided. For that reason, we used the Me1 binding sites predicted by the data generator, which can be downloaded from the following website: <http://www.bcgsc.ca/downloads/histone/human/HeLa/H3K4me1/stimulated/>

2.2 De novo motif discovery for sequences around STAT1 binding sites

2.2.1 Obtaining DNA sequences around high confidence STAT1 binding sites

We selected the top 10,000 STAT1 binding sites for analysis, that's because we wan to focus on these high confidence binding sites, also, we want to save computation time.

For the selected STAT1 binding sites, we obtained 401bp DNA sequences around each of them (i.e., 200bp upstream and 200bp downstream). That

2.2. *De novo motif discovery for sequences around STAT1 binding sites*

is, we composed a bed file containing chromosome, start location and end location for each binding site, then load the bed file to UCSC database and obtained the corresponding DNA sequences (Kent et al. [30]).

2.2.2 Handling the repeating regions

Repeats are typically masked out for ChIP-chip experiment simply because they occur too often in the genome, and one could never tell where within the repeat region is being bound by a TF. This was standard Affymetrix design at an earlier time, and even now on tiled arrays. Even with sequencing approaches, most people do not deal with repeats for the same reason. When obtaining the DNA sequences, we also tend to avoid the repeats.

When obtaining sequences from UCSC Genome Browser, we let the sequence which overlaps with simple tandem repeats in more than 50% of its regions be filtered out. By doing this, we obtained 9,992 401bp STAT1 sequences for the top 10,000 STAT1 binding sites.

For the 9,992 sequences obtained, we further masked out the simple tandem repeats within them using DecoyMasker, which is one of the packages provided by CREAD project (Smith et al. [62]). The settings we used for DecoyMasker are in the appendix.

2.2.3 De novo motif discovery with GADEM

Reason of choosing GADEM for de novo motif discovery

GADEM can be viewed as an extension of MEME, and the major difference between them is the way they initialize the starting PWM: MEME uses all the subsequences as potential starting PWM, and after running EM for one iteration, the subsequence giving largest likelihood is chosen as starting PWM; GADEM initializes the PWM using spaced dyad, which is two over represented k-mers separated by spacers, and that is a more efficient way for initialization.

The large amount of sequences we need to handle is the main reason for choosing GADEM: a comparison of GADEM and several well known methods, such as GAME, MEME and Weeder, showed that GADEM can detect motifs much faster than the other methods. What is more, in a simulation study, the motifs GADEM predicted was at competitive accuracy (Li [39]).

Parameter setting in GADEM

The settings are in the appendix.

Here are some important parameter settings: 1) We set the parameter minN to 200; this restriction let the binding site corresponding to each predicted motif occur at least 200 times in the sequences (which is 5% of the total number of STAT1 sequences). 2) We set p-value for declaring a subsequence as motif to 0.0002. The way GADEM calculates p-value is: it transforms PWM into an integer score matrix, the exact score distribution is then determined for a fixed-length subsequence (i.e. each possible score corresponding to a fixed-length subsequence can be calculated given the score matrix), as discussed by Staden (Staden et al. [64]). With the score distribution, the p-value of a specific sequence can be easily obtained.

We noticed that GADEM purposely allows site overlapping to avoid a site being assigned to a unique motif, and that leads to overlapping issue where two motifs are assigned to the same location. In order to keep the proportion of two motifs' overlapping locations in a reasonable range, we tuned the parameter SIMILARITY_ALPHA in defines.h from 0.35 to 0.40. After getting de novo predicted motifs, we checked the overlapping of predicted binding sites between each pair of motifs. We found that overlapping exists, but there were less than 30% of overlapping between the sites of any two motifs. We did not combine the overlapping binding sites, as the overlapping may be of biological importance, instead of being the "useless side product". This will be further discussed in the result part.

2.2.4 Visualization of de novo predicted motifs

De novo discovered motifs in PWM form

GADEM represent a de novo discovered motif with position weight matrix (PWM). In a PWM, each column represent a position in the motif, and each entry of the matrix is the occurrence rate of A, C, G and T at a specific position.

For example, the following matrix is a PWM of STAT1 GAS motif. The matrix shows that the motif consists of 11 nucleotides. It also shows that 52% of the binding sites declared to have this motif have nucleotide T in the first position, and 48% of them have other nucleotides in the first position; more than 96% of the binding sites declared to have this motif have nucleotide T in the second position; more than 94% of the the binding sites declared to have this motif have nucleotide T in the third position; and so on. We can tell that in this motif, the second and the third positions are

2.2. De novo motif discovery for sequences around STAT1 binding sites

more conserved than the first position, because nucleotide T occur most of the time at second and third position.

> STAT1.TTTCyrGGAAA

A 0.064 0.009 0.011 0.081 0.048 0.441 0.019 0.010 0.966 0.971 0.501

C 0.208 0.019 0.013 0.874 0.593 0.012 0.002 0.004 0.015 0.008 0.045

G 0.205 0.006 0.027 0.014 0.009 0.352 0.970 0.973 0.011 0.013 0.227

T 0.523 0.966 0.949 0.032 0.350 0.195 0.009 0.014 0.007 0.009 0.226

Using WebLogo to visualize de novo predicted motif

PWM of a motif can be visualized in a way that is easier for human to recognize it and tell the sequence component and conservation at each position within it.

For example, with WebLogo application (Crooks et al [12]), a motif is represented with a series of stacks of nucleotides: the overall height of one stack indicates the motif conservation at the corresponding position, which is the difference between maximum possible entropy and the entropy of the observed symbol distribution.

Entropy of a discrete variable X is calculated in the following way:

$$H(X) = E(I(X)) \quad (2.1)$$

Where $I(X)$ is the information content of X: $I(x_i) = -\log_2 p(x_i)$.

For a position within a DNA motif, the maximum possible entropy is reached when the occurrence rate of each one of the four nucleotides equals to $\frac{1}{4}$:

$$\sum_4 \frac{1}{4} (-\log_2 \frac{1}{4}) = 2 \quad (2.2)$$

The entropy of the observed symbol distribution for a position can be calculated as:

$$\sum_4 p_n (-\log_2 p_n) \quad (2.3)$$

Here, p_n is the probability that a position have nucleotide n. And $\sum p_n = 1$.

In the WebLogo, the height of a position is calculated as:

$$\sum_4 \frac{1}{4} (-\log_2 \frac{1}{4}) - \sum_4 p_n (-\log_2 p_n) = 2 - \sum_4 p_n (-\log_2 p_n) \quad (2.4)$$



Figure 2.1: WebLogo representation of STAT1

Therefore, in the WebLogo representation, the maximum height of a position is 2 bits when probability of observing a specific type of nucleotide in that position is 1, and probability of observing the other three type of nucleotide in that position is 0; and the minimum height of a position is 0 bit when $p=0.25$ for each type of possible nucleotide observed in that position. The higher a stack, the more conserved the motif is in the corresponding position of that stack; and the height of symbols within the stack indicates the relative frequency of each nucleic acid at that position. Figure 2.1 is the WebLogo visualization of STAT1 GAS motif corresponding to the PWM we showed previous text. We can tell from the logo that the GAS motif is composed of 11 nucleotides, it is well conserved at position 2, 3, 4, 7, 8, 9 and 10.

2.3 Obtaining gene information data set

We obtained the gene information data set from UCSC website. UCSC known gene is for protein coding genes and each know gene is substantiated by at least a transcript record (either a GenBank mRNA or a NCBI RefSeq) and a UniProt protein record (Hsu et al. [26], Kent et al. [30]).

We chose to use UCSC known gene information for our analysis because: 1) it has larger coverage of human genome compared with other gene information, such as RefSeq from NCBI, Ensembl Genes from EMBL-EBI; 2) The genes are substantiated by both transcript record and protein record, which made the information reliable. From UCSC website, we downloaded two tables, knownGene.txt and kgXref.txt, for Human genome18:

<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/>

The most useful information is: 1) a gene's TSS is in the 4th column of knownGene.txt (if it is coded on forward strand) or in the 5th column of knownGene.txt (if it is coded on reverse strand). 2) a gene's gene symbol is in the 5th column of table kgXref.txt. 3) knownGene.txt and kgXref.txt are related with each other by the gene name column.

For some genes, their chromosome names are followed by _random or _hap. 'random' is for unassembled chromosome, and 'hap' is for haplotype chromosome. We did not use those genes in the analysis. Also, many genes have the same TSS, we did not use duplicated promoter regions in the analysis.

2.4 Relating STAT1 binding with DE genes

2.4.1 Genes differentially expressed in IFN-gamma stimulated HeLa cell

In the work of Hartman et al., 63 genes that show differential expression (DE) after the IFN-gamma stimulation were identified on Chromosome 22 of HeLa cell . We used the DE genes they reported (Hartman et al. [24]).

2.4.2 Genes differentially expressed in other three types of IFN-gamma stimulated human cells

Microarray data sets for different types of human cells stimulated by IFN-gamma

We searched for IFN-gamma stimulated human cell time series microarray data from GEO web site. We got microarray data set for other three types of human cells, which was carried out for all the known human genes. The name of data sets are listed below:

microarray dataset 1: Anti-IFN antibody 16 array set (Peripheral Blood Mononuclear Cell)

microarray dataset 2: Interferon gamma effect on keratinocytes: time course (skin cell)

microarray dataset 3: IFN-gamma-inducible gene expression in Toxoplasma-infected human fibroblasts (fibroblast cell)

We decided to get differentially expressed DE genes ourselves because we could not get the txt format file of the DE genes reported in the papers.

2.4. Relating STAT1 binding with DE genes

Also, when selecting the DE genes ourselves, we applied same FDR criteria for the three data sets, which made the results consistent and comparable.

Handling the missing data

We used R package EMV to handle the missing data, which used k-nearest neighbor method (we set the nearest neighbor as 5). EMV can be downloaded from CRAN project webpage:

<http://cran.r-project.org/src/contrib/Archive/EMV/>

Microarray1 and microarray3 data sets are cDNA microarray experiments, and they both have experiments at 4 time points. Microarray1 has one experiment in each time point, while microarray3 has two replicated experiments at each time point. For these two data sets, we selected the genes whose expression value is available in 3 out of 4 (or 6 out of 8) experiments. We then used knn method to fix the missing values, setting the nearest neighbour as 5.

Microarray2 data set is from Affymetrix platform, which is carried out in 4 time points after IFN-gamma stimulation. At each time point, the mRNA level of stimulated cell and control cell was detected with separate arrays. Totally, there are 8 arrays. We selected the genes who has at least 6 expression value in the 8 arrays (i.e., genes with ABS_CALL as Present). We used the microarray data as it was, and did not use KNN to fix the value whose ABS_CALL is Absent or Marginal.

Get the differentially expressed genes

For each microarray dataset, we used an R package, EBarrays, to detect the differentially expressed genes, and controlled the FDR at 0.05. Here, a DE gene is a gene having mRNA level variation in at least one time point out of the four time points.

2.4.3 Checking STAT1 binding site in promoter region of DE genes

We used the files knownGene.txt to relate the location of STAT1 binding sites and promoter of DE genes. As the TSS and the direction of gene coding is known, we obtained -4,000~+2,000bp region of each TSS as the promoter region. For each promoter region, we checked whether there is STAT1 binding site in it.

2.5 Multiple test for proportions

We have several sets of samples, and for each sample, we know the sample size and proportion of the sample with a specific character. We use multiple test for proportion to decide whether there is significant difference between proportions from each pair of samples.

2.5.1 Testing proportion of two samples

Suppose that we have two sets of samples, X and Y, and we know that the sample sizes are n_x and n_y , each x and y either has a specific character or does not have it.

We know that $\hat{p}_x = \frac{\sum x_i}{n_x}$, where p_x is proportion of samples in X having the specific character. Our sample size and the number of sample with/ without specific character is large enough (i.e., number of STAT1 sequences with/ without motif_x is always greater than 50). Therefore according to central limit theorem, \hat{p}_x approaches normal distribution with mean equaling to p_x and variance equaling to $\frac{p_x(1-p_x)}{n_x}$. We can make the same inference for \hat{p}_y . Therefore, we can test whether p_x is significantly different from p_y with z test, as described in the book by Simonoff (Simonoff [61]).

2.5.2 Multiple test with Bonferroni adjustment

As we have several samples and need to do test for each pair of samples here, we adjusted the p-values for multiple testing using a Bonferonni type control (Simes [60]).

Another thing to mention is that the pooled estimate of the common proportion (\bar{p}) is used for every pair-wise hypothesis test: $\bar{p} = \frac{c_1+c_2+\dots+c_g}{n_1+n_2+\dots+n_g}$, where g is total number of samples we have, c_i is the number of individuals from sample_i with the specific character.

Chapter 3

Results

3.1 Analyzing ChIP-Seq data

3.1.1 Prediction of STAT1, Pol2 binding sites from ChIP-Seq data

Several software packages are available for analyzing ChIP-Seq data. Initially, we wanted to use PICS to analyze the ChIP-Seq data of STAT1 and Pol2. However, at that time PICS was still under development, and it did not give result fast enough.

We used MACS (Zhang et al. [74]) to analyze STAT1 and Pol2 ChIP-Seq data with control data. We chose to use MACS for two main reasons: 1) MACS is easy to use and gives out result in fairly short time. 2) MACS performs better than other peak-finding software as for FDR and the motif occurrence within 50 bp of the peak centers, as described in the paper by Zhang et al.

The settings for MACS that we used are specified in the appendix.

Table 3.1 shows the number of ChIP-Seq reads, the number of control reads and the predicted binding sites for STAT1 binding sites and Pol2 binding sites.

In order to focus our analysis on the high confidence STAT1 binding sites and save the computational time, we selected the top 10,000 STAT1 binding sites (according to FDR) for further analysis.

3.1.2 Prediction of Me1 flanked regions of biological interest

We obtained the predicted Me1 sites for IFN-gamma stimulated HeLa cell that was provided by Dr. Robertson (Robertson et al. [54]).

Heintzman et al. reported that there was Me1 occurring in a large proportion of active promoters and active enhancers. Moreover, they reported the bimodal distribution of Me1 centered around the TSS (Heintzman et al. [25]).

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

Table 3.1: Summary of ChIP-Seq reads and binding site predicted for STAT1 and Pol2

	STAT1	Pol2
Number of reads from ChIP-Seq experiment	26,731,492	29,060,928
Number of reads from control experiment	23,435,631	29,840,987
Number of binding sites predicted	24,751	33,067

Based on their finding, we checked the distance between each of two nearby Me1 sites (Figure 3.1 shows the distance distribution). If the distance between two Me1 sites was in the range of 200~1,000bp, we recorded the positions of these two Me1 sites as the ends of a biologically meaningful Me1 flanked region. We decided that the Me1 flanked region is a good indication of active promoter or enhancer (i.e., promoter or enhancer for a gene actively transcribed).

In total, we recorded 90,113 Me1 flanked regions as biologically meaningful flanks from 301,493 predicted Me1 binding sites, and used them in the following analysis.

3.2 Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

STAT1, Pol2 binding and Me1 flanked regions are important factors related to the transcription of genes.

We think that is interesting to know how these three factors are located with respect to the transcription start site (TSS). For example, it is interesting to know whether they all tend to occur in the upstream of a TSS instead of the downstream of a TSS, and whether they are all located very close to the TSS.

Here, we analyzed the length of the upstream or downstream region of TSS versus the number of non-redundant regions with at least one STAT1 binding site. We did a similar analysis for the Pol2 binding sites and Me1 flanked regions.

Also, we checked the intersection of locations of these three factors in the whole genome. The work by Robertson et al. showed that large proportion of STAT1 binding sites were related to Me1 binding sites (Robertson et al.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

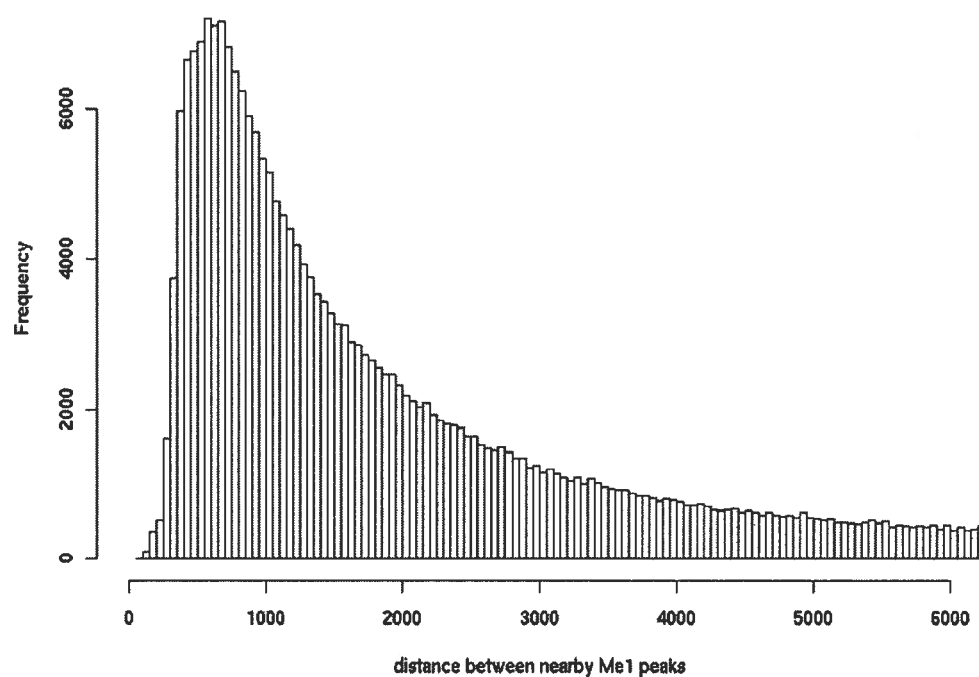


Figure 3.1: Distribution of the distance between nearby Me1 binding sites.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

[54]). Therefore we expected that the locations of many STAT1 binding sites would intersect with Pol2 binding sites and Me1 flanked regions.

3.2.1 Relationship between the length of region upstream or downstream of TSS and the number of regions having STAT1, Pol2 binding site and Me1 flanked region

We used all non-redundant TSS (42645 TSS in total), and got their up or downstream regions with lengths of 200, 400, 600, ... , 40,000bp. Then we checked the total number of regions with at least one STAT1 binding site detected by ChIP-Seq.

We found that the increase in the length of a region upstream or downstream of TSS leads to the increase in the number of regions with at least one STAT1 binding site, which indicated that TF binding is not restricted to the regions around the TSS, instead, TF occurs every where in the genome. Also, at the same length, the number of upstream regions with STAT1 binding site is always greater than the number of downstream regions with STAT1 binding site. Result is shown in Figure 3.2.

We did similar analysis for Pol2 binding sites and Me1 flanked regions (For each Me1 flanked region, we used the center of the region as its location). Result is shown in Figure 3.3 and 3.4.

Figure 3.2 shows that at the same length, there are more upstream regions than downstream regions have STAT1 binding site. This phenomenon can be explained by STAT1's function: as a TF, STAT1 regulates the gene's expression by binding to its upstream region.

Figure 3.3 shows that at the same length, more downstream regions than upstream regions have Pol2 binding site. As a comparison, Rozowsky et al. had aggregated the short reads of Pol2 over regions proximal to TSS of all consensus coding sequences human genes, and their analysis showed that Pol2 reads from ChIP-Seq occur more often in upstream regions of TSS than in downstream regions of TSS(Rozowsky et al. [57]). It seemed that our finding on Pol2 binding sites was not consistent with their finding on Pol2 reads. Initially, we thought this may be caused by the control reads highly enriched in upstream of TSS. Therefore, we ran MACS on Pol2 ChIP-Seq data without giving it the control data, and then checked the Pol2 binding sites' occurrence in the upstream or downstream regions again. In this way, we found that when the region length is shorter than 15,000bp, there are always more downstream regions with Pol2 binding sites than the upstream regions with Pol2 binding sites; when the region length is greater

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

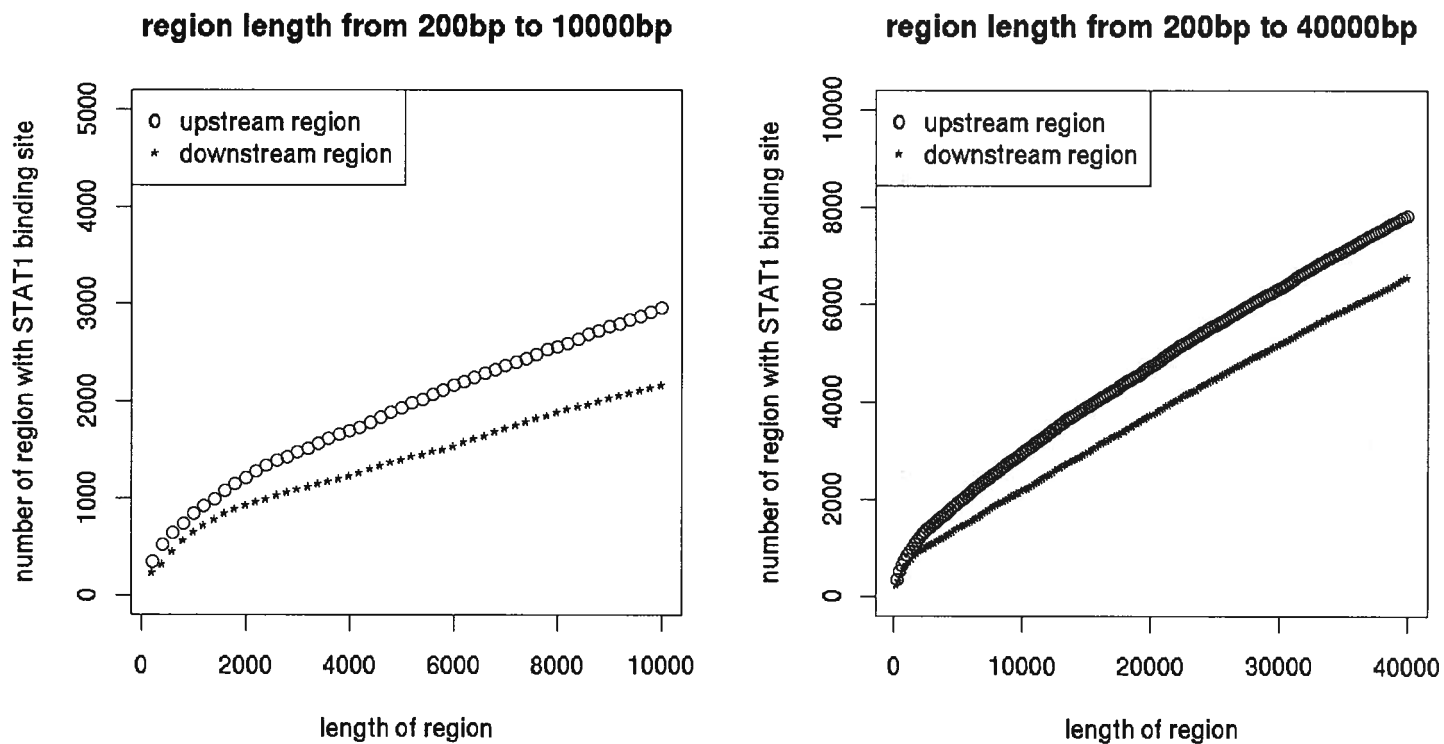


Figure 3.2: Relationship between length of upstream or downstream region of TSS and the number of regions with at least one STAT1 binding site. On the left, the upstream or downstream regions of length between 200bp and 10,000bp. On the right, the upstream or downstream regions of length between 200bp and 40,000bp.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

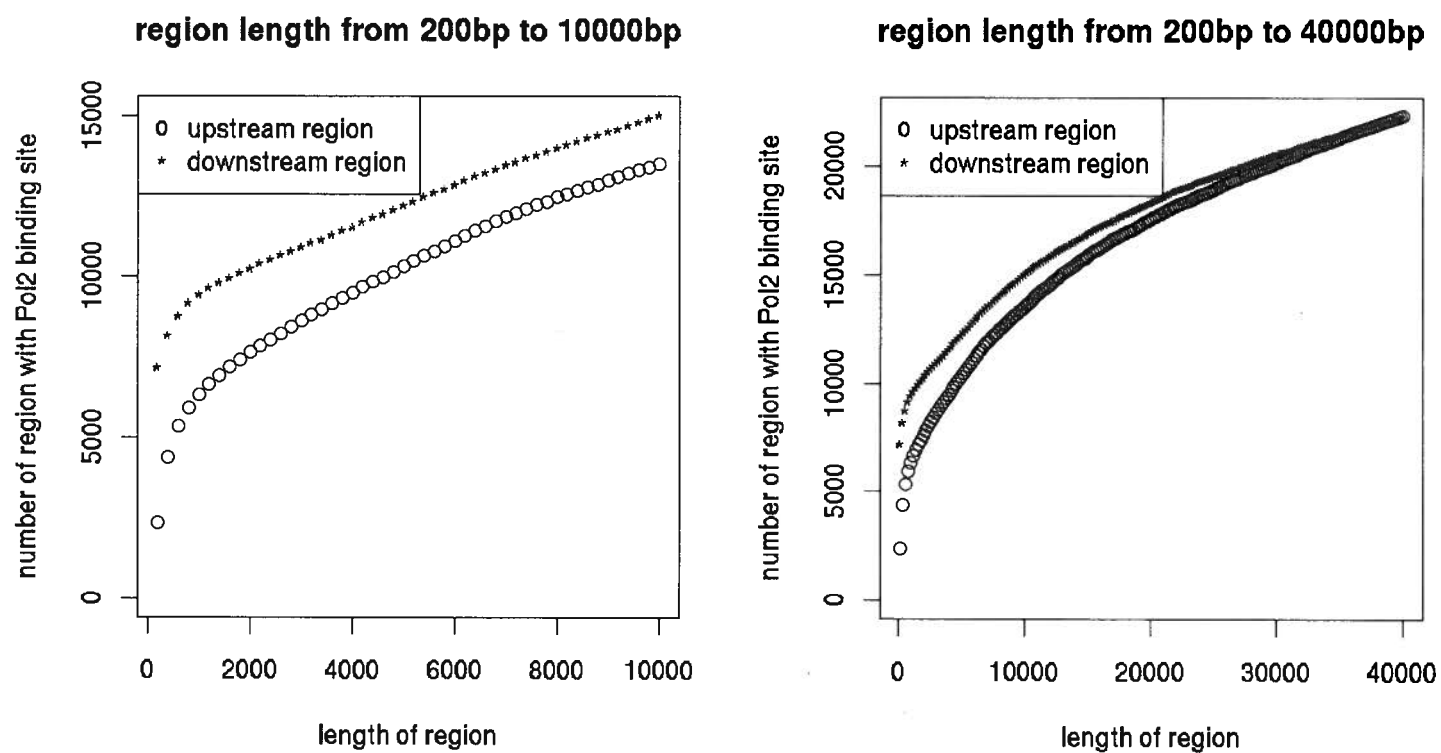


Figure 3.3: Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Pol2 binding site. On the left, the upstream or downstream regions of length between 200bp and 10,000bp. On the right, the upstream or downstream regions of length between 200bp and 40,000bp.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

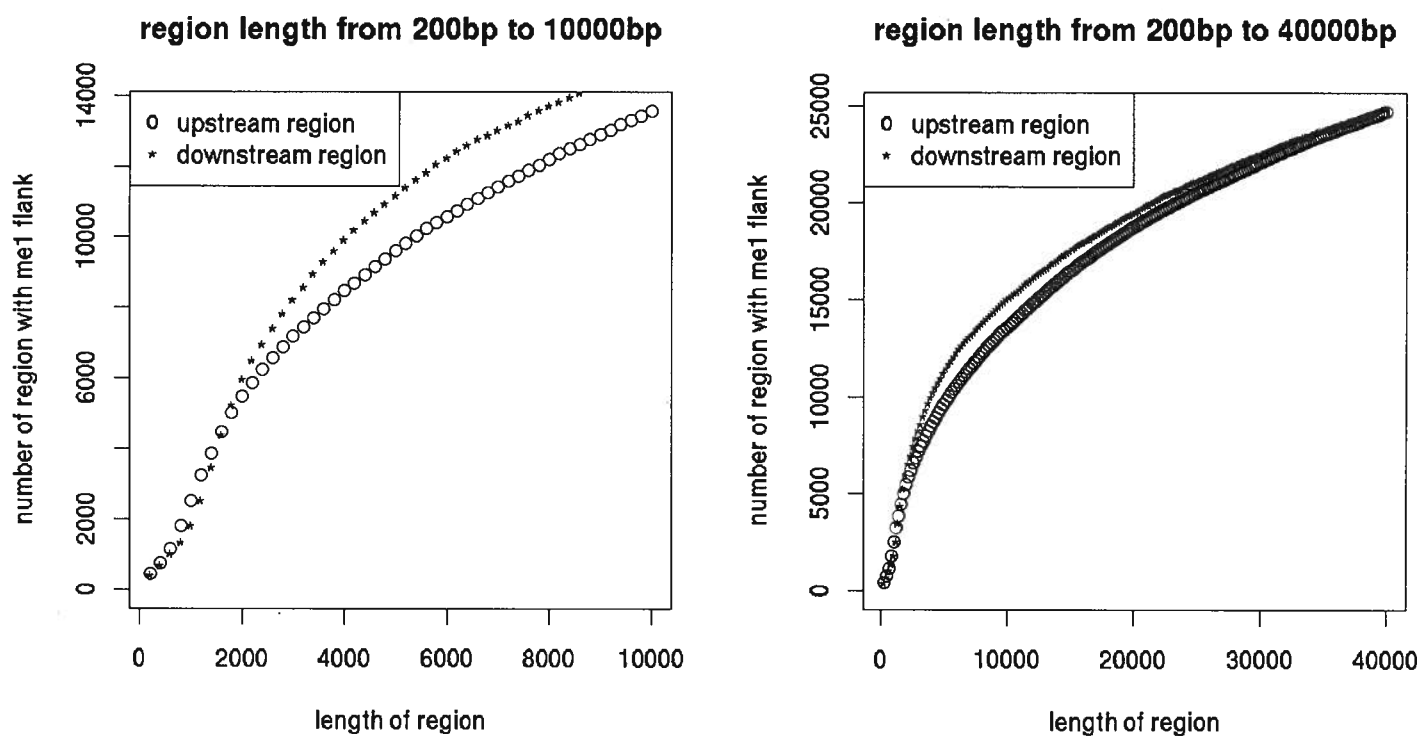


Figure 3.4: Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Me1 flanked region. On the left, the upstream or downstream regions of length between 200bp and 10,000bp. On the right, the upstream or downstream regions of length between 200bp and 40,000bp.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

than 15,000bp, there are always more upstream regions with Pol2 binding sites than the downstream regions with Pol2 binding sites. The result is shown in Figure 3.5.

We think a possible explanation for the inconsistency of our finding for Pol2 binding sites and Rozowsky's finding for Pol2 reads regarding the TSS is: There are lots of reads that are mapped back to the upstream regions of TSS, however, for many upstream regions, the density of reads are not intense enough to be called as peaks by MACS. In contrast, the total number of reads that are mapped back to the downstream of TSS is not that large, but for many downstream regions, the density of these reads are intense enough to be called as peaks.

Furthermore, we know that Pol2 synthesize RNA by binding to a gene's promoter region and sliding along the gene, which explains our finding well. Another two possible explanations for why Pol2 binding sites occur more frequently in downstream region are: 1) Pol2 stalling occur more often in the downstream of TSS than in the upstream of TSS (Zeitlinger et al. [72]). 2) There are many unknown TSS (an alternative promoter) downstream of the reported TSS where Pol2 can bind to.

Figure 3.4 shows that at the same length, there are more downstream regions than upstream regions which have Me1 flanked region. We are not sure how to explain this phenomenon. Maybe biologists can help to explain this phenomenon.

3.2.2 The inconsistency in the increase in number of regions having STAT1, Pol2 binding sites and Me1 flanked regions with respect to the increase in the length of regions

In our study, the increase in the length of a region is 200bp at each interval, while the increase in the number of regions with STAT1 binding site is not at a consistent rate. We plotted the increase in the number of regions with STAT1 with respect to the region length. The result is shown in Figure 3.6. A similar analysis was done for Pol2 binding sites and Me1 flanked regions, which is shown in Figure 3.7 and 3.8.

Figure 3.6 shows that, for the downstream region, as the region length increases, the number of regions with STAT1 also increases; when the region length is less than 2,000bp, the increase in the number of regions with STAT1 binding site is always greater than 50 as the region length increases 200bp; when the region length is more than 2,000bp, the increase in the number of regions with STAT1 binding sites is always less than 50 as the region

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

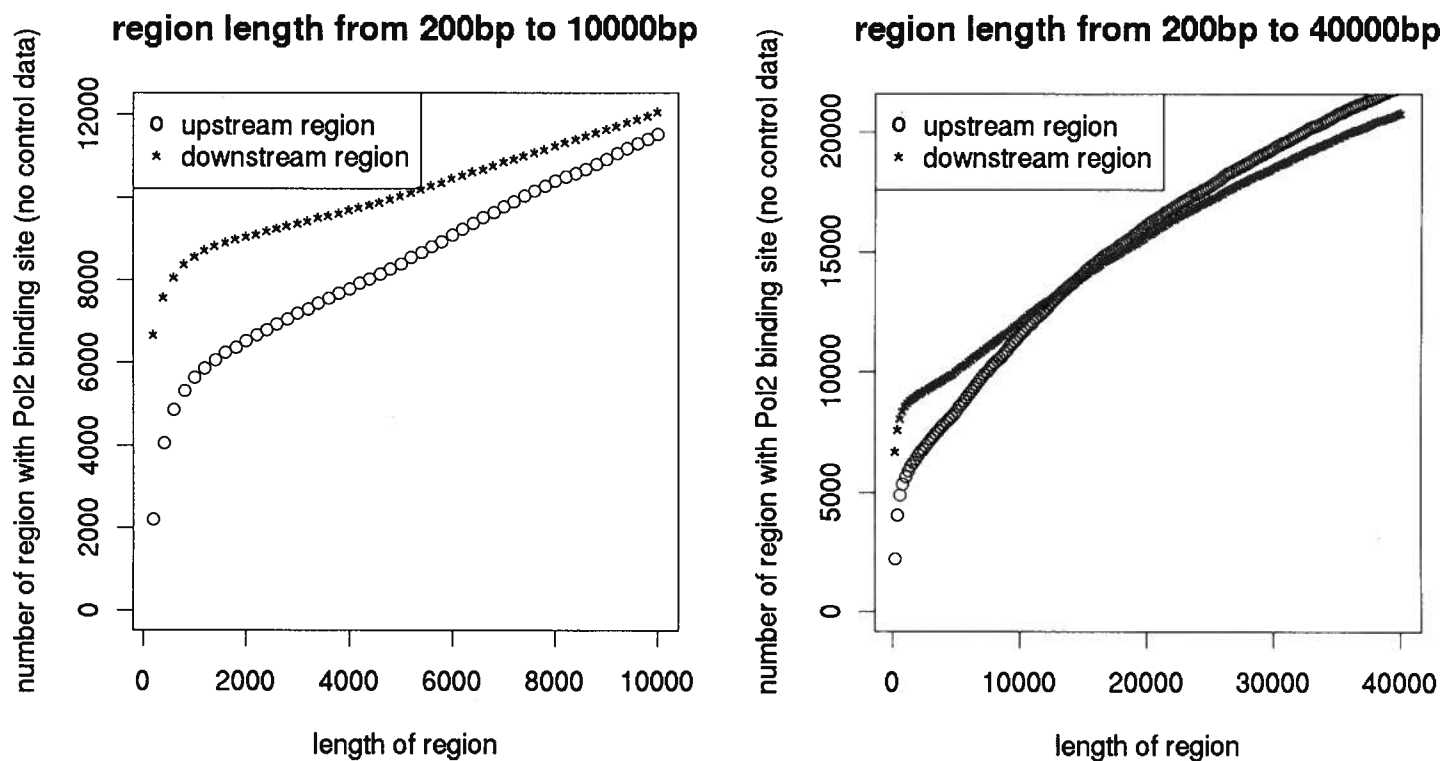


Figure 3.5: Relationship between length of upstream or downstream region of TSS and the number of regions with at least one Pol2 binding site detected without control data. On the left, the upstream or downstream regions of length between 200bp and 10,000bp. On the right, the upstream or downstream regions of length between 200bp and 40,000bp.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

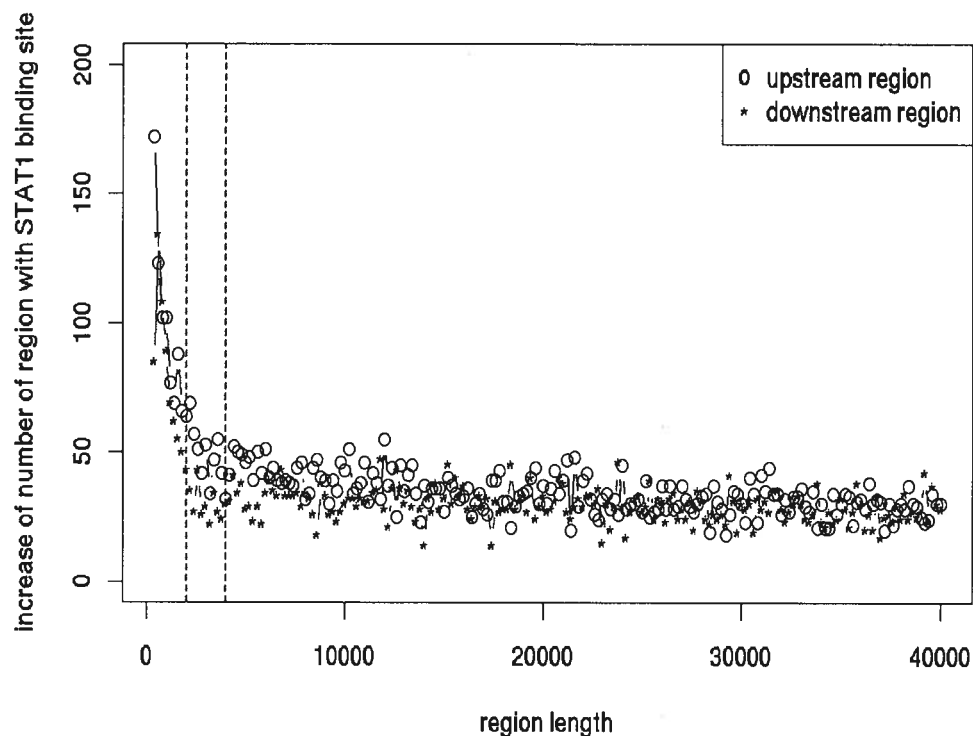


Figure 3.6: Increase of number of upstream or downstream region of TSS with STAT1 with respect to the region length; the region length increases by 200bp for each pair of neighbouring points (i.e., the first data point shows the increase in the number of regions having STAT1 binding site as the region length increases from 200bp to 400bp). Green line indicates the region length at 2,000bp, blue line indicates the region length at 4,000bp.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

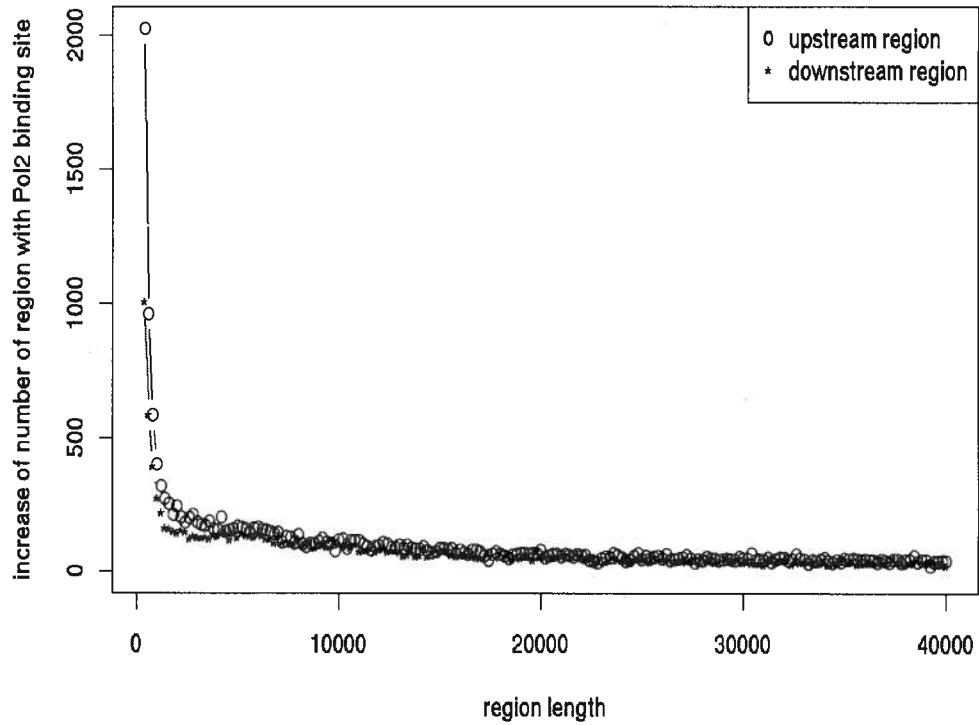


Figure 3.7: Increase of number of upstream or downstream region of TSS with Pol2 with respect to the region length; the region length increases by 200bp for each pair of neighboring points (i.e., the first data point shows the increase in the number of regions having Pol2 binding site as the region length increases from 200bp to 400bp).

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

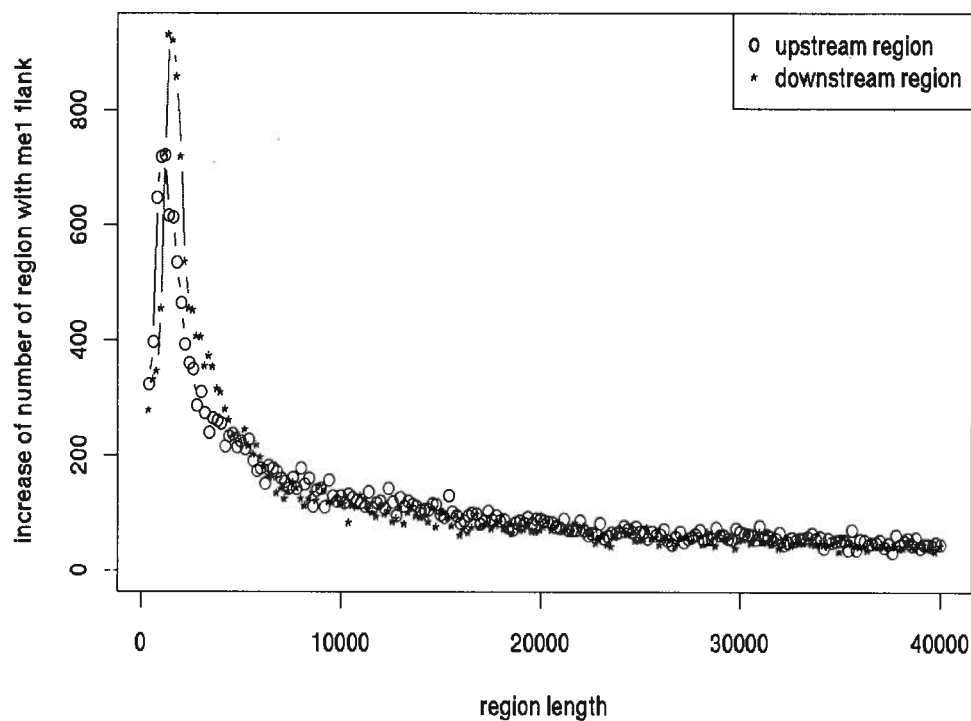


Figure 3.8: Increase of number of upstream or downstream region of TSS with Me1 flanked regions with respect to the region length; the region length increases by 200bp for each pair of neighboring points (i.e., the first data point shows the increase in the number of regions having Me1 flanked region as the region length increases from 200bp to 400bp).

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

length increases 200bp. For the upstream region, the boundary appears at 4,000bp (in the figure, 2,000bp and 4,000bp are indicated with green line and blue line separately). We conclude that STAT1 binding sites are located everywhere in the genome, and many of them are located near TSS.

From Figures 3.7 and 3.8, we got a similar conclusion for Pol2 binding sites and Me1 flanked regions: they are located everywhere on the genome, and they occur more often in regions close to TSS than in regions distant from TSS.

What is more, Figure 3.8 shows that as the region length increases, the number of upstream or downstream regions with Me1 flanked regions increases, and the rate of the increase in number of regions turns faster at first and turns slower later. It is known that Me1 occurs in enhancer regions as well as at promoter regions. That explains why the rate turned faster at first for the upstream region. Yet, it is hard to explain why the rate turns faster at first for the downstream region. Maybe this phenomenon can be explained with further investigation on Me1.

3.2.3 Intersection of locations of STAT1 binding sites, Pol2 binding sites and Me1 flanked regions

We checked the intersection of location of STAT1, Pol2 binding sites and Me1 flanked regions.

Here we define that, a STAT1 binding site and a Pol2 binding site have intersection if the distance between them is less than 2,500 bp; a STAT1/Pol2 binding site and Me1 flanked region have intersection if the STAT1/Pol2 binding site locates in a 200~1,000bp Me1 flanked region; a STAT1 binding site, a Pol2 binding site and a Me1 flanked region have intersection if both the STAT1 and the Pol2 binding site are located within a Me1 flanked region (in this case, the distance between the STAT1 binding site and the Pol2 binding site is shorter than 1,000 bp).

A Venn diagram for the intersection of STAT1 binding sites, all Pol2 binding sites and all Me1 flanked regions is shown in Figure 3.9.

We found that there are intersections for these three factors. Especially, there are 370 Me1 flanked regions which have both STAT1 and Pol2 binding sites located within them. We think that STAT1 in these regions is likely to actively regulate the gene transcription. Further analysis for these STAT1 binding sites showed that 137 of them are located in 6,001bp promoter regions (4,000bp upstream and 2,000bp downstream) of genes; as a comparison, 2,307 of the total 9,992 STAT1 binding sites are located within the promoter regions. We found that the proportion of the STAT1 in pro-

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

STAT1 peaks, Pol2 peaks and Me1 flanks

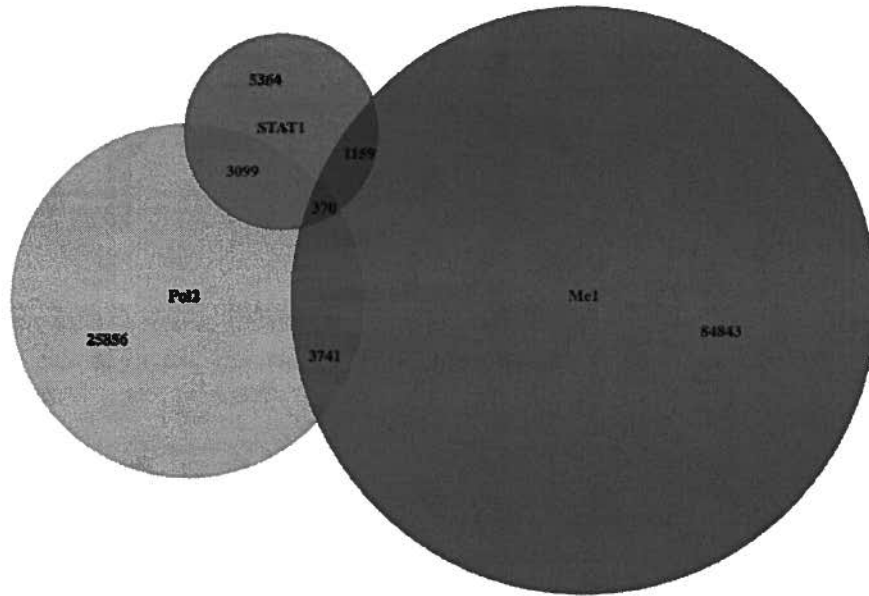


Figure 3.9: Intersection of STAT1 binding sites, all Pol2 binding sites and all Me1 flanked regions. Out of the 9,992 STAT1 binding sites, 5,364 of them are not in the Me1 flanked regions and far from Pol2 binding sites, 3,469 of them are close to Pol2 binding sites, 1,529 of them are in Me1 flanked regions, and 370 of them are in Me1 flanked regions and close to Pol2 binding sites.

moter region is higher for the STAT1 belonging to the 370 regions than for all the STAT1, with p-value less than 0.05. The p-value was calculated as described in the method section.

There are large number of STAT1 binding sites far from Pol2 binding site and not in Me1 flanked region. Three possible reasons can explain this phenomenon: 1) These STAT1 binding do not cause Pol2 binding or Me1 flanked region, i.e., they did not change the DNA character with respect to Po2 and Me1. 2) A lot of Pol2 binding sites and Me1 binding sites have not been detected through ChIP-Seq experiments or the analysis on ChIP-Seq data. 3) The Pol2 experiment was carried out for unstimulated HeLa cell. It is possible that if the Pol2 binding sites are detected for the IFN-gamma stimulated HeLa cell, the intersection of Pol2 and STAT1 will be larger.

3.2. Relationship between STAT1, Pol2 binding site, Me1 flanked region and TSS

STAT1, top 10K Pol2 and top 20K Me1

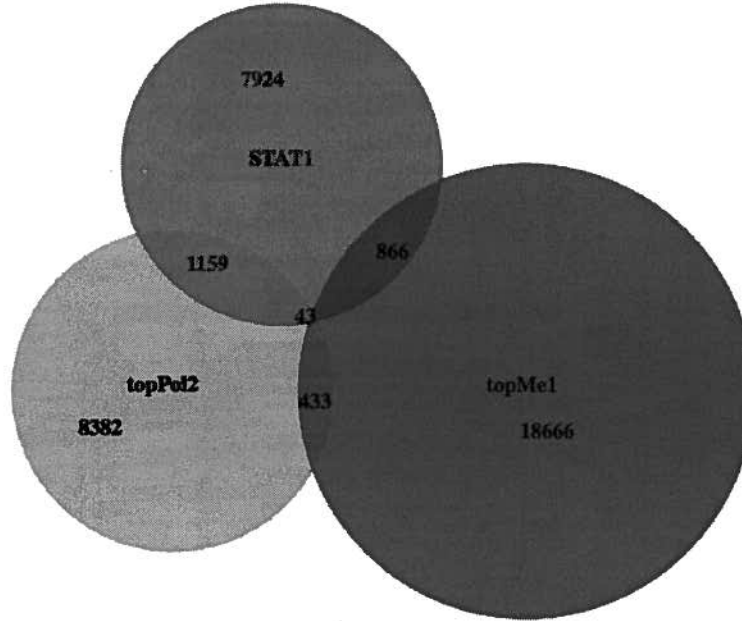


Figure 3.10: Intersection of STAT1 binding sites, top 10K Pol2 binding sites and top 20K Me1 flanked regions.

We then selected 10,000 high confidence Pol2 binding sites according to the FDR predicted by MACS; and got the top ~20,000 Me1 flanked regions for high confidence Me1 binding sites which have large number of reads in them. We checked the intersection of location of STAT1 binding sites, top Pol2 binding sites and top Me1 flanked regions again. The result is shown in Figure 3.10.

Comparing Figure 3.9 and Figure 3.10, we found that as the number of Pol2 binding sites selected decreased from 33,066 to 10,017, the number of STAT1 binding sites which are close to Pol2 binding site decreased from 3,469 to 1,202 ($33,055/10,017$ is slightly larger than $3,469/1,202$); as the number of Me1 flanked regions selected decreased from 90,113 to 20,008, the number of STAT1 binding sites which are in Me1 flanked region decreased from 1,529 to 909 ($90,113/20,008$ is much larger than $1,529/909$).

The intersection between Pol2 and STAT1 decreased a little slower than

the decrease of the number of Pol2 binding sites; while the intersection between Me1 and STAT1 decreased much slower than the decrease of the number of Me1 flanked regions. The Venn diagram indicates that we may have some false positive Pol2 binding sites, and we may have more false positive Me1 flanked regions: if there is no false positive discovery in Pol2 or Me1, we should expect that as the number of Pol2/Me1 selected decreases, the intersection between Pol2 or Me1 and STAT1 should decrease at the same rate. Therefore, it is necessary to find out and use binding sites with higher confidences in the future.

3.3 De novo motif discovery for STAT1 sequences

Here, we predicted which TFs may collaborate with STAT1 in regulating the gene transcription through de novo motif discovery for sequences around STAT1 binding sites. To do this, we utilized the information of STAT1 binding sites predicted through ChIP-Seq data, and the information of the human genome sequence.

For 401bp (i.e., 200bp upstream and 200bp downstream) regions flanking 9,992 STAT1 binding sites, we got 13 de novo motifs predicted by GADEM. We will collectively refer to the de novo discovered motifs as motif_x (x=1,2,3,...,13), and their corresponding TFs as "TF_x".

3.3.1 Frequency of the binding sites corresponding to de novo discovered motifs occurring in the STAT1 sequences

For each de novo discovered motif, GADEM predicts the occurrence of its corresponding binding site in the STAT1 sequences.

We checked the number of times that binding site of each motif occurring in the 9,992 STAT1 sequences. The result is shown in Figure 3.11.

As Figure 3.11 shows, most motifs' binding sites occur in more than 4,000 out of the 9,992 STAT1-centered sequences. It indicates that the binding sites of de novo discovered motifs co-occur very often, and that these binding sites may form cis-regulatory module in promoter regions of genes.

We noticed that binding site of a motif sometimes occur more than once in a sequence. For example, binding site of STAT1 occurs in 6,678 of the 9,992 STAT1 sequences, and the binding site occur 8,940 times in total.

The proportion of the STAT1 sequences with binding sites corresponding to the STAT1 motif (6,678 out of 9,992) is not high. There are two main

3.3. De novo motif discovery for STAT1 sequences

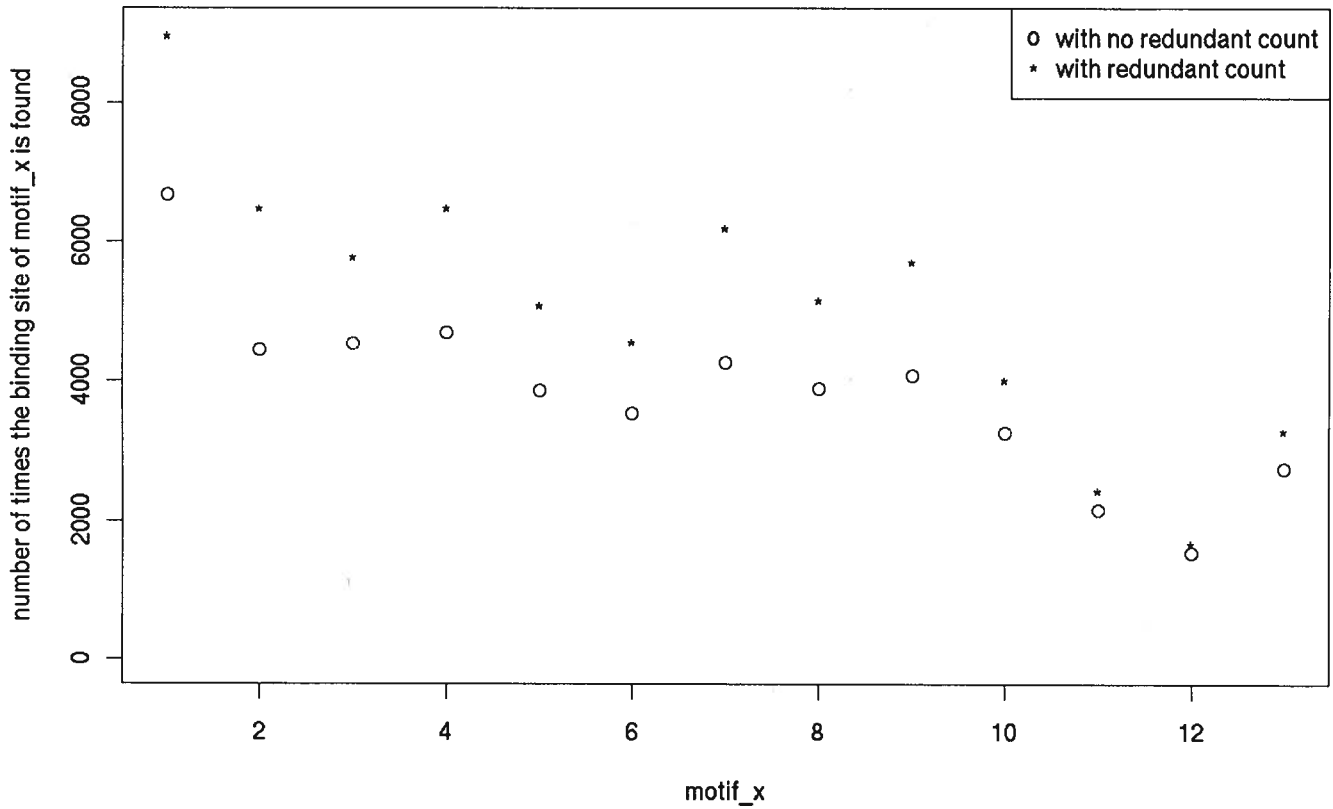


Figure 3.11: Number of times the binding sites of de novo motifs are found in 9,992 401-bp STAT1 sequences. We obtained the occurrence time with or without redundant counts: a binding site may occur more than once in a sequence. “with redundant” means that we count the exact times that a binding site occur; “without redundant” means that we count the number of sequences having at least one binding site of motif_x. The five de novo predicted motifs selected for in-depth analysis are highlighted in red.

reasons which can explain this result. Firstly, there was noise in the prediction of STAT1 binding sites through analyzing ChIP-Seq data. Secondly, the STAT1 may not bind to the DNA sequence of the detected sites directly i.e., it may bind to Pol2 or other TF which binds to the DNA sequence.

3.3.2 Visualizing the de novo discovered motifs

We used WebLogo for visualizing de novo discovered motifs and their most similar counterpart in STAMP. We also plotted the histogram for the distribution of locations of the de novo motifs in the 401-bp regions flanking STAT1 binding sites. Results are in Table 3.2.

Motif_x and its most similar counterpart in STAMP

For each motif_x, we looked for the known motif which is most similar to it: for a given motif, STAMP compares it with all the known motifs in database, and gives the top ranking matches; at the same time, STAMP provides e-value which indicates the similarity between the query motif and the corresponding motif in database: more similar two motifs are, smaller the e-value is.

In the first column of Table 3.2, we put motif_x's label. In the second column of the table, we put each motif_x's WebLogo on top of the WebLogo of its corresponding STAMP motif, and put e-value at the bottom.

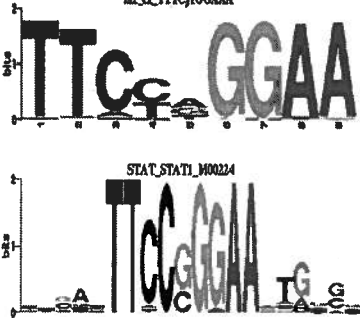
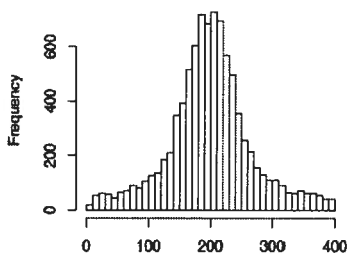

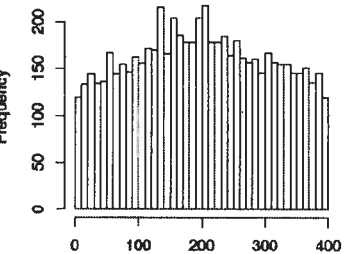

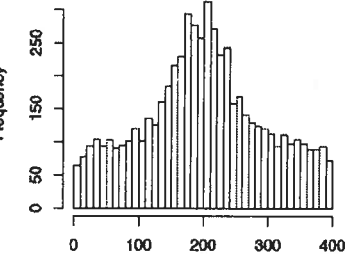
The distribution for the locations of binding sites of de novo discovered motifs

We checked the distribution for the locations of each motif's binding site in the 401bp sequence around the STAT1 binding site identified: we recorded the location of a motif in the 401bp STAT1 sequence if the motif is predicted to be in that region, and we represented the distribution for the location of a motif's binding site in all the sequences with histogram. The histogram is in the third column of Table 3.2.

In the histogram, the peak of the distribution indicates the region where the motif_x occurs most frequently. A high peak in the center of the 401bp region indicates that binding site corresponding to motif_x occur close to the STAT1 binding site most of the time. The motif's location is an important factor for selecting motifs which are most likely to cooperate with STAT1 in the following analysis.

3.3. De novo motif discovery for STAT1 sequences

Table 3.2: De novo motifs predicted for 401bp sequences around the STAT1 binding sites.

Motif_x predicted by GADEM	logo of motif_x logo of motif_x's counter part in STAMP	location distribution of motif_x
m1	<p>ml.cl_TTTCyGGAAA</p>  <p>STAT_STAT1_M00224</p> <p>e-value: 7.6653e-08</p>	
m2	<p>m2.cl_nCCGACGn_rev</p>  <p>CH1_MAZ_M00648</p> <p>e-value: 3.5317e-07</p>	
m3	<p>m3.cl_4GGAAAAG_rev</p>  <p>homer_Nasag_M01123</p> <p>e-value: 2.6355e-07</p>	

Continued on the next page

3.3. De novo motif discovery for STAT1 sequences

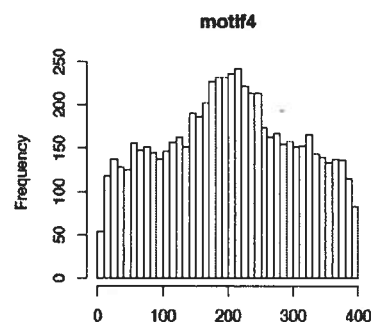
Table 3.2 – continued from the previous page

Motif_x predicted by GADEM logo of motif_x location distribution of motif_x
 logo of motif_x's counter part
 in STAMP

m4



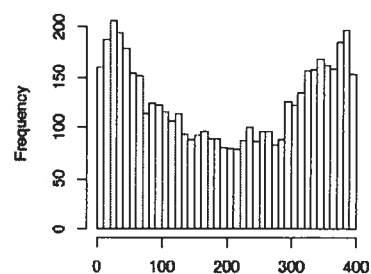
not similar to any STAMP motif
 with e-value less than $1e-6$



m5



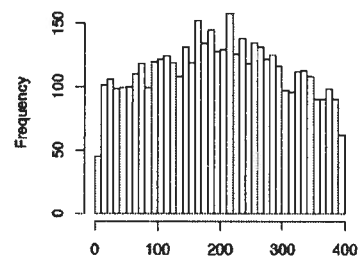
e-value: $9.9695e-08$



m6




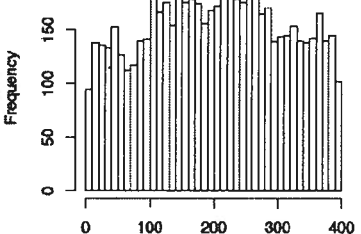

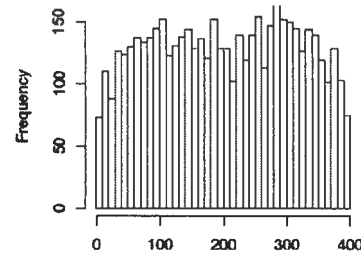
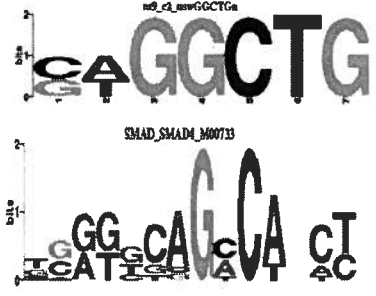
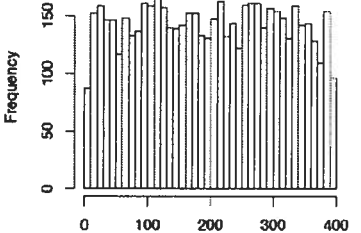
e-value: $9.9089e-11$



Continued on the next page

3.3. De novo motif discovery for STAT1 sequences


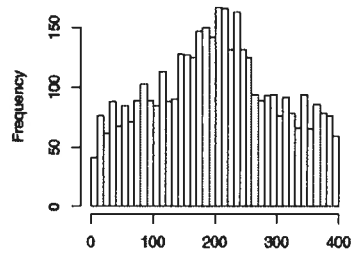

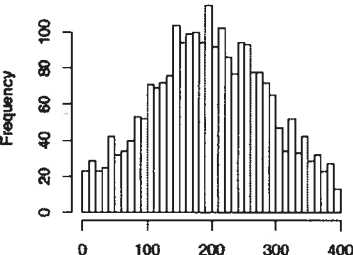

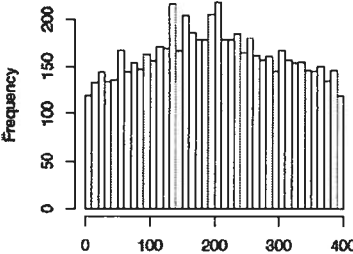
Table 3.2 – continued from the previous page

Motif_x predicted by GADEM	logo of motif_x logo of motif_x's counter part in STAMP	location distribution of motif_x
m7	 <p>not similar to any STAMP motif with e-value less than 1e-6</p>	
m8	 <p>not similar to any STAMP motif with e-value less than 1e-6</p>	
m9	 <p>e-value: 1.2077e-06</p>	

Continued on the next page


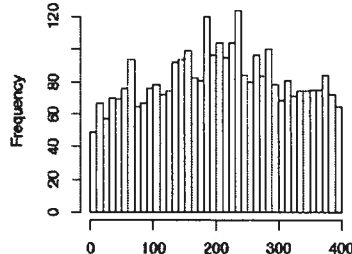
3.3. De novo motif discovery for STAT1 sequences

Table 3.2 – continued from the previous page

Motif_x predicted by GADEM	logo of motif_x logo of motif_x's counter part in STAMP	location distribution of motif_x
m10	<p>  </p> <p>e-value: 4.0197e-10</p>	
m11	<p>  </p> <p>e-value: 3.4770e-11</p>	
m12	<p>  </p> <p>e-value: 6.7164e-06</p>	

Continued on the next page

Table 3.2 – continued from the previous page

Motif_x predicted by GADEM	logo of motif_x logo of motif_x's counter part in STAMP	location distribution of motif_x
m13	 <p>e-value: 1.6392e-08</p>	

3.3.3 Location overlapping issue of binding sites corresponding to de novo detected motifs

We had 13 de novo motifs in total. For each pair of the motifs, we checked the frequency of their binding site locations overlapping with each other (here, we define two motif locations as overlapping if their centre are within 5bp in a sequence). Result is shown in Figure 3.12.

Figure 3.12 shows that the pair of motif1&motif3 and the pair of motif1&motif4 have more than 1,000 overlapping locations. In the STAT1-centered sequences, motif1 occurs 8,940 times; motif3 occurs 5,772 times; motif4 occurs 6,476 times.

As Table 3.2 shows, the PWM of motif1, motif3 and motif4 are similar to each other to some extent. Figure 3.14 illustrates a region of Chromosome 1, where the locations of motif1, motif3 and motif4 overlaps.

We are not sure whether the overlapping locations are a side-product of motif-discovery tool (i.e., they correspond to the same motif, but GADEM identified them as separate motifs because of the initialization), or whether it has a biological meaning.

It has been reported that two TFs bind to their binding sites whose locations overlap with each other, which makes it more convenient for the TFs to interact with each other. For example, Ganster et al. identified a region with NF- κ B and STAT1 binding sites overlapping with each other in the -5.8Kb promoter region of inducible nitric oxide synthase gene (iNOS)

3.3. De novo motif discovery for STAT1 sequences

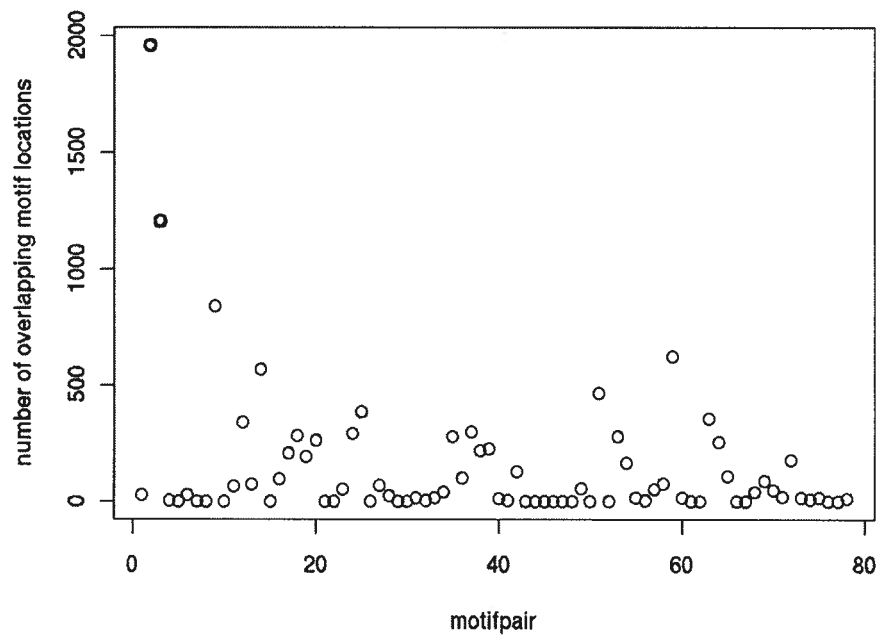


Figure 3.12: Number of times the location of two motifs overlapping with each other. (In total, we have 13 motifs, and 78 ($13 \cdot 12 / 2$) motif pairs). The pairs, motif1&motif3 and motif1&motif4, are highlighted in red.

3.4. Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

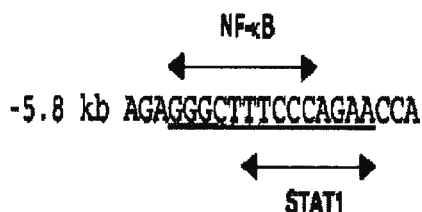


Figure 3.13: The overlapping of the STAT1 binding site and the NF-κB binding site in the genomic region of -5.8kb in iNOS promoter (as reported by Ganster et al. [19]).

(Figure 3.13). They conducted a mutation experiment on iNOS promoter-reporter plasmid, which is -7.2kb of the upstream DNA of iNOS linked to the luciferase reporter gene. They reported that mutations in both the NF-κB and STAT1 binding sites at -5.8kb completely eliminated the expression of cytokine-induced luciferase, while mutation of either NF-κB or STAT1 sites individually failed to inhibit the promoter activity. Also, their result of gel shift analysis suggested that the region with NF-κB and STAT1 motif is bifunctional, and can be bound by both NF-κB and STAT1 (Ganster et al. [19]). Another example is given by Kang et al. (Kang et al. [29]), who found that there is Egr-1 motif, and potential motifs of YY1 and SP1 in the TGF-responsive region of Id1 promoter, and the motifs of Egr-1, YY1 and SP1 are overlapping with each other.

We believe that a genome-wide ChIP-chip or ChIP-Seq experiment on motif_x will help to answer our question about the biological significance of overlapping.

3.4 Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

By comparing the PWM with motifs of known genes, we know that motif1 is STAT1 GAS motif. It is known that STAT1 activates or represses gene transcription primarily by binding to GAS motif as a homodimer, and it also binds to interferon-stimulated response elements (ISRE motif). In our de novo motif prediction, only GAS motif was obtained.

For the remaining 12 de novo discovered motif_x, we filtered out the motifs which look like a noise, and kept motifs whose corresponding TFs are

3.4. Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

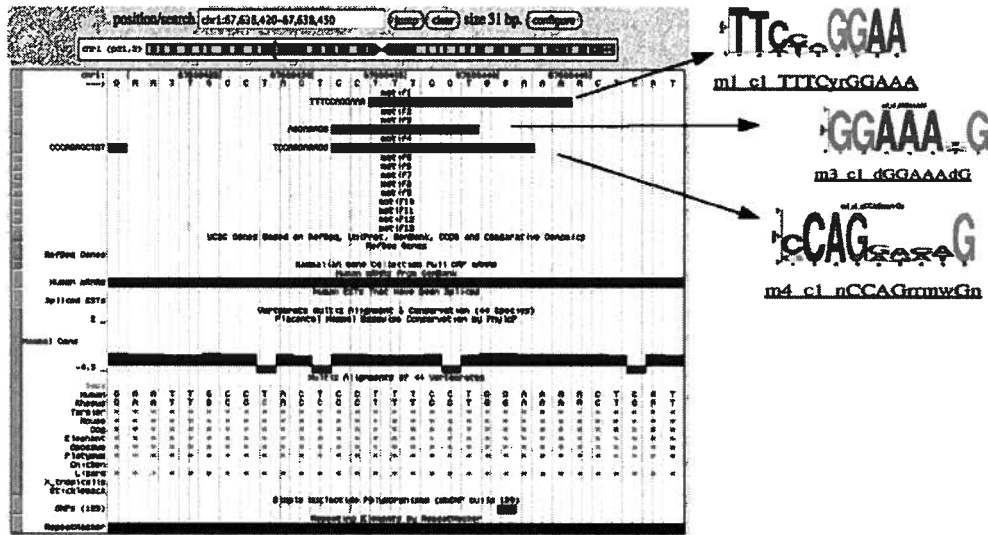


Figure 3.14: An example of three motifs whose predicted binding sites are overlapping with each other.

most likely to cooperate with STAT1. In this way, we kept the number of motifs for further inspection small, so that we can explore in more detail for the selected motifs.

The filtering was done according to the motif's component and location distribution. Basically, we filtered out a motif if it has repeating nucleotides or if it is not very conserved in most positions; Also, we want to get the motifs which are always located very close to STAT1. If the distribution of a motif's location does not have a high peak in the middle of the 401bp STAT1 binding site, we decided that it does not always occur very close to STAT1 binding site, and filtered it out.

After filtering, we obtained 5 motifs (including STAT1 GAS motif), which are listed in Table 3.3. Note that:

- in the first column, we put the label of motif_x, the potential name(s) of motif_x (we list names of all STAMP motifs which are similar to motif_x with e-value less than $1e^{-6}$, ordered by similarity between STAMP motif and motif_x). Below each name, we list whether there is evidence (from the literature or the data set) showing the co-occurrence of motif_x's STAMP match and STAT binding site, i.e., the cooperation of the two TFs.



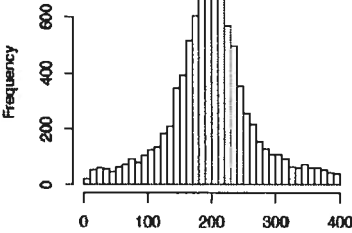


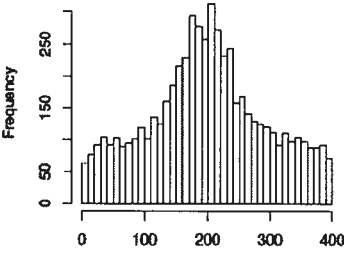
- in the second column, we put the logo of de novo predicted motif on the top, and put its corresponding STAMP match(s) beneath it (all the matches

3.4. Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

with e-value less than $1e^{-6}$ are shown).

- in the third column, we put the distribution of locations of binding site corresponding to motif_x, in the -200~+200bp sequence flanking the identified STAT1 binding site.




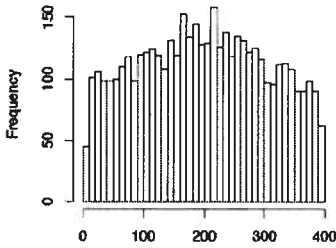
Table 3.3: Five selected de novo motifs predicted for 401bp sequences around the STAT1 binding sites.

motif_x and its most similar counterpart(s) in STAMP (with evidence supporting the cooperation of motif_x's counterpart(s) and STAT1 if available)	<div> <div>logo of motif_x</div> <div>logo of motif_x's counterpart(s) in STAMP</div> <div>location distribution of motif_x</div> </div>
<p>m1</p> <p>STAT1: Although the motif is most similar to STAT5 in STAMP, we decide that it should be STAT1. Because the ChIP-Seq experiment was done for STAT1.</p>	<div>    </div> <p>e-value: 7.6653e-08</p>
<p>m3</p> <p>Nanog: cooperation of Nanog and STAT1 reported in the literature</p>	<div>    </div> <p>e-value: 2.6355e-07</p>

Continued on the next page

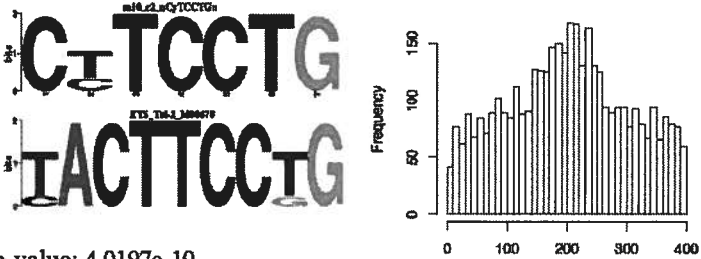

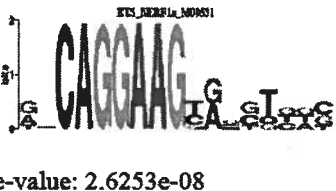


3.4. Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

Table 3.3 – continued from the previous page

<div> <div>m6</div> <div> <div>HEB: evidence for the cooperation of HEB and STAT1 via analyzing HEB ChIP-chip data and STAT1 ChIP-Seq data</div> <div>AP4: no evidence for the cooperation of AP4 and STAT1</div> </div> </div>	<div> <div> <div>m6_HZ1_CCGCGCGG</div>  </div> <div> <div>m6_HZ1_CCGCGCGG</div>  </div> <div>e-value:9.9089e-11</div> </div> <div> <div> <div>m6_HZ1_CCGCGCGG</div>  </div> <div>e-value: 9.8993e-07</div> </div> <div>  </div>
Continued on the next page	

3.4. Selection of motif_x whose corresponding TFs are most likely to cooperate with STAT1

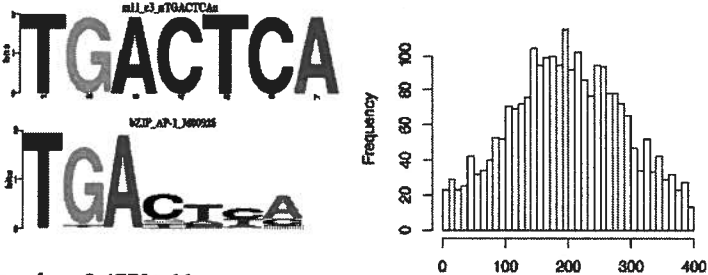


Table 3.3 – continued from the previous page

<p>m10 Tel2: no evidence for the cooperation of Tel2 and STAT1 NERF1, Ets2, Ets1 (TFs from ETS family): cooperation of TFs from ETS family with STAT1 reported in the literature</p>	<div data-bbox="690 394 1386 651">  <p>e-value: 4.0197e-10</p> </div> <div data-bbox="690 672 1023 861">  <p>e-value: 2.2596e-08</p> </div> <div data-bbox="690 903 1023 1092">  <p>e-value: 2.6253e-08</p> </div> <div data-bbox="690 1123 1023 1312">  <p>e-value: 3.8687e-08</p> </div> <div data-bbox="690 1354 1023 1549">  <p>e-value: 1.2329e-07</p> </div>
--	---

Continued on the next page

3.5. Literature review for the cooperation of STAT1 and selected motif_x

Table 3.3 – continued from the previous page

<p>m11 AP1: cooperation of AP1 and STAT1 reported in the literature Bach2 and GCN4 (TFs from bZIP family): no evidence for the cooperation of Bach2 or GCN4 with STAT1</p>	 <p>e-value: 3.4770e-11</p>
	 <p>e-value: 3.9058e-10</p>
	 <p>e-value: 7.8747e-10</p>

3.5 Literature review for the cooperation of STAT1 and selected motif_x

3.5.1 Cooperation of STAT1 and Nanog

Sun et al. compared the expressions of human and mouse genes which are critical in the pathways for embryonic stem cell differentiation, and they got cross-species conserved co-expression gene clusters.

Through analyzing the promoters of the conserved co-expression genes, they found that STAT, Nanog, and several other TFs had binding sites in most of the promoter of co-expressed genes in all the examined pathways. They concluded that these TFs conduct key regulatory mechanisms underlying the evolutionary conserved co-expression (Sun et al. [65]).

3.5. Literature review for the cooperation of STAT1 and selected motif_x

Table 3.4: High confidence HEB regions with STAT1 GAS binding site on Chromosome 19

HEB region's start on Chr19	HEB region's end on Chr19	in a promoter region or not
1083120	1083513	in promoter of ZNF181
5938429	5939109	not in promoter region
39916349	39916987	in promoter of SBNO2
40619367	40619805	not in promoter region
45801040	45801577	not in promoter region
49403136	49403820	not in promoter region

3.5.2 Cooperation of STAT1 and HEB

Through searching GEO database, we found that there is ChIP-chip done for HEB on human chromosome 19 (Gardini et al. [18]).

Gardini et al. predicted 1,023 HEB binding site regions on Chromosome 19 of human cell (U937) in normal condition, and the regions of binding sites are summarized in table S9 of their paper.

Among the 9,992 STAT1 binding sites gained from ChIP-Seq experiment, 283 of them were located in Chromosome 19.

We found that 21 HEB binding site regions have at least one STAT binding site within them.

Furthermore, we required that the HEB binding site region being used should actually have a subsequence like HEB motif (we used Motiflocator (Thijs et al. [68]) to check whether a region has HEB motif.); we also required STAT1 binding site used to have a subsequence like STAT1 GAS motif in the -200~+200bp region around it (GADEM result is used to judge whether there is subsequence like STAT1 GAS motif). With that stringent criteria, we found 6 HEB binding regions have at least one STAT1 binding site within them, and we refer to them as high confidence intersection, which is shown in Table 3.4. Two of the six HEB binding regions are located in -4,000~+2,000 promoter regions.

Our findings indicate that there are binding sites of HEB and STAT1 located close to each other on Chromosome 19, and these two TFs may cooperate in regulating the genes.

3.5.3 Cooperation of STAT1 and Tel2 (or TFs from the ETS family)

Tel2 is the motif the most similar to de novo predicted motif₁₀. We searched for the cooperation between STAT1 and Tel2, but did not find any literature regarding the subject.

We noticed that motif₁₀ is also similar to several TFs from ETS family. Therefore, we searched for the literature discussing the cooperation of STAT1 and TFs from the ETS family.

STAT1 and Ets-1 binding site in promoter of *bcl-x*

Fuijo et al. carried out electrophoretic mobility shift assay with SIE as probe for the promoter region of *bcl-x* gene. They found that the formation of SIE-STAT1 complex was inhibited by oligonucleotide containing the GAS motif, and they concluded that there is a GAS motif of STAT1 at -41bp of the *bcl-x* gene. Moreover, they constructed *bcl-x* promoter-luciferase reporter plasmid by linking -161 10bp promoter region of human *bcl-x* and luciferase gene, then conducted mutagenesis analysis. The analysis showed, mutation of the GAS motif will result in the reduction of the promoter activity under LIF stimulation (Fuijo et al. [17]).

The alignment of mouse and human *bcl-x* gene promoter region showed that there is consensus Ets-1 binding site in -425bp -437bp of human *bcl-x* promoter and Ets-1 binding site in -419bp -431bp of mouse *bcl-x* promoter (Grillot et al. [23]).

STAT1 and Ets-2 binding sites in promoter of ICAM-1

Launoit et al. conducted transient transfection assays on human ICAM-1 gene, and found that two Ets proteins, Ets-2 and ERM significantly activate the transcription of ICAM-1 promoter. With electrophoresis assay and DNase footprinting, they identified two Ets binding sites at positions -158 and -138 from the TSS of ICAM-1 (Launoit et al. [38]).

Duff et al. showed that pervanadate treatment of human cell stimulates the protein complex formation on pI γ RE motif (located -76~-66bp of the ICAM-1 TSS), and that the complex contains STAT1. The treatment also induced the activation of the ICAM-1 gene (Duff et al. [14]).

3.5.4 Cooperation of STAT1 and AP1

We found several literatures indicating the cooperation of STAT1 and AP1.

3.5. Literature review for the cooperation of STAT1 and selected motif_x

STAT1 and AP1 binding site in iNOS promoter

Ganster et al. used gel shift assay for the -5.2kb region of the human iNOS promoter, they found that IFN-gamma or cytokin mixture induced a protein-DNA complex; mutation of the STAT1 site abolished the protein-DNA binding. Using a supershift assay with antibody for STAT1, they also confirmed that STAT1 binds to the DNA at -5.2kb in the iNOS promoter (Ganster et al. [19]).

Kristof et al reported that there are two AP-1 motifs in the -5155 -5131bp human iNOS promoter, which is bound by heterodimer. Removal of the two AP-1 sites decreased iNOS's response to LPS or IFN-gama stimulation 3.5 fold. They concluded that activation of the human iNOS promoter by cytokines (i.e., tumor necrosis factor-alpha , interleukin-1beta, IFN-gamma) required downstream and upstream AP-1 transcription factor binding sites (Kristof et al. [34]).

Interestingly, there are both STAT1 and AP1 binding sites in the promoter of iNOS in the mouse cell: Gao et al reported that iNOS gene expresses when the mouse macrophage cell is stimulated by IFN-gamma or lipopolysaccharide (LPS). The binding of STAT1 to the iNOS promoter's GAS site is necessary for the expression of iNOS gene induced by IFN-gamma or LPS (Gao et al. [20]). Lowenstein et al found that with minimal promoter construct, luciferase reporter gene expressed little when the cells were stimulated by LPS or IFN-gamma. While the expression of luciferase reporter gene increased notably under the stimulation of LPS or IFN-gamma when the NOS 5' flanking region, which contains motif of AP1, was placed upstream of the gene (Lowenstein et al. [41]).

STAT1 and AP1 binding site in VIP promoter

Researchers found that 1,330bp upstream of the TSS of vasoactive intestinal peptide gene (VIP), there is cytokine response element (CyRE), which contains STAT and AP-1 binding sites. Symes et al. found that in one of the regions within the CyRE, cytokine treatment induces binding of a protein complex composed of the members of STAT transcription factor family (STAT1 α and STAT3). Mutation of this STAT-binding site attenuates cytokine-mediated transcriptional activation. And activation of STAT transcription factors contributes to the induction of the VIP gene(Symes et al. [66]). In another experiment, they constructed luciferase reporter plasmid, which is 180bp CyRE linked to the luciferase reporter gene. They found that mutation in the AP-1 proteins did not bind to the CyRE with mutated AP-1

3.6. Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

binding site, and the mutation of AP1 site reduced the CNTF-mediated induction of luciferase by 50% compared with the reporter plasmid with wild type CyRE (Symes et al. [67]).

STAT1 and AP1 binding site in *bcl-x* promoter

As described before, there is STAT1 binding site in the promoter region of *bcl-x* (Fuijo et al. [17]).

The alignment of mouse and human *bcl-x* gene promoter region showed a consensus AP1 binding site in -270bp -260bp of human *bcl-x* promoter and AP1 binding site in -266bp -256bp of mouse *bcl-x* promoter (Grillot et al. [23]).

As the literature review indicates, there are STAT1, Ets-1 and AP1 binding site in the promoter of *bcl-x*, and these three binding sites are located close to each other. In our de novo prediction, we also got some STAT1 sequences having these three motifs predicted within them. We think that binding sites of STAT1, Ets-1 and AP1 may form cis-regulatory element in some promoters and regulate the gene expression together.

STAT1 and AP1 binding site in *beta*-defensin-2 promoter

Mineshiba et al. reported that there are subsequences like tandem STAT binding site and AP-1 binding site in the promoter of human *beta*-defensin-2 (hBD-2). They suspected that the STAT binding site may play a role in the regulation of the promoter activity (Mineshiba et al. [46]).

Kanda et al. reported recently that antisense oligonucleotides against AP-1 components suppresses hBD-2 production; antisense oligonucleotide against STAT1 also suppressed hBD-2 production (Kanda et al. [28]).

3.6 Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

Inspired by the work of Kielbasa et al. (Kielbasa et al. [31]), we checked whether genes potentially regulated by STAT1 and other TF_x (TFs corresponding to motif_x are collectively referred to as TF_x) participate in the same biological process. We believe that genes potentially regulated by STAT1 and other TF_x participating in the same biological process can provide further evidence to support our prediction that STAT1 may cooperate with TF_x in regulating the gene transcription level.

3.6. Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

Table 3.5: Total number of genes in each group of genes potentially regulated by STAT1 and TF_x

Group	Number of genes in the group
genes potentially regulated by STAT1 and Nanog	3,745
genes potentially regulated by STAT1 and HEB	3,111
genes potentially regulated by STAT1 and Tel2	3,783
genes potentially regulated by STAT1 and AP1	1,431

We first extracted 1,000bp upstream and 200bp downstream sequence of the non-redundant TSS, and used these sequences as the promoter regions of genes.

After getting the PWMs of STAT1 GAS motif and other 4 motif_x (Nanog, HEB, Tel2 and AP1) predicted by GADEM, we used Cluster-Buster to identify whether there is cluster of STAT1 GAS motif and one of the four motif_x in a gene's promoter region (with default setting). In that way, we got 4 groups of genes which have cluster of binding sites corresponding to STAT1 GAS motif and motif_x in their promoter regions, and the genes in a group are potentially regulated by STAT1 and one of the TF_x.

The number of genes in each group is listed in Table 3.5.

For each of the 4 groups, we used GOrilla to test whether the genes belonging to it have enriched Gene Ontology (GO) term(s) of biological process, compared with all the genes as background. We found: a), 3 groups have enriched GO biological process terms with p-value less than $1e^{-9}$; b), the groups of genes potentially regulated by different combinations of STAT1 and motif_x have different GO enrichment terms. The Gene Ontology highlighting the biological processes that are significantly enriched in genes potentially regulated by STAT1&Nanog, STAT1&HEB, STAT1&Tel2, STAT1&AP1 are shown in Figure 3.16, Figure 3.17, Figure 3.18, and Figure 3.19 respectively. Note that in the GOrilla output, each node in the graph is a biological process term; the darker the color the smaller the p-value, and the more significantly enriched a biological process term is (the color and corresponding p-value is shown in Figure 3.15).

We applied same method on human genes by giving Cluster-Buster STAT1 PWM only. The result is shown in Figure 3.20.

Figure 3.20 shows there are some enriched biological process for the genes potentially regulated by STAT1, but there is no biological process

3.6. Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

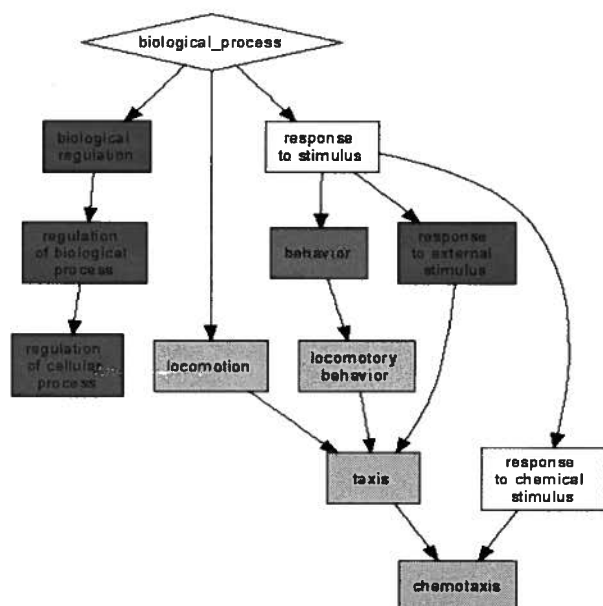


Figure 3.17: Gene Ontology highlighting the biological process significantly enriched in genes which are potentially regulated by STAT1 and HEB.

with p-value less than $1e^{-9}$. This could be due to the fact that STAT1 is a fairly general TF, and it can regulate transcription of large number of genes, which are involved in diverse biological processes.

We admit that it is possible for a gene which is not regulated by a TF to have subsequence like motif of that TF in its promoter region, and there could be many false positive for the genes predicted to be regulated by STAT1. We also know that in eukaryotic cell, the regulation of gene is typically achieved by binding of several TFs onto its promoter region. Therefore, in a gene's promoter region, if there are two or more subsequences that are similar to motifs of TFs and are located close to each other, the probability that the gene is regulated by these TFs is larger. Comparing Figures 3.16, 3.17, 3.18, 3.19 and 3.20, we found many genes with combination of motifs for STAT1&Nanog or STAT1&HEB or STAT1&Tel2 in the promoter regions participate in several specific biological processes. We infer that STAT1 cooperates with Nanog/HEB/Tel2 in regulating genes' expression, and binding sites of STAT1 and Nanog/HEB/Tel2 are likely to be in cis-regulatory element.

3.6. Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

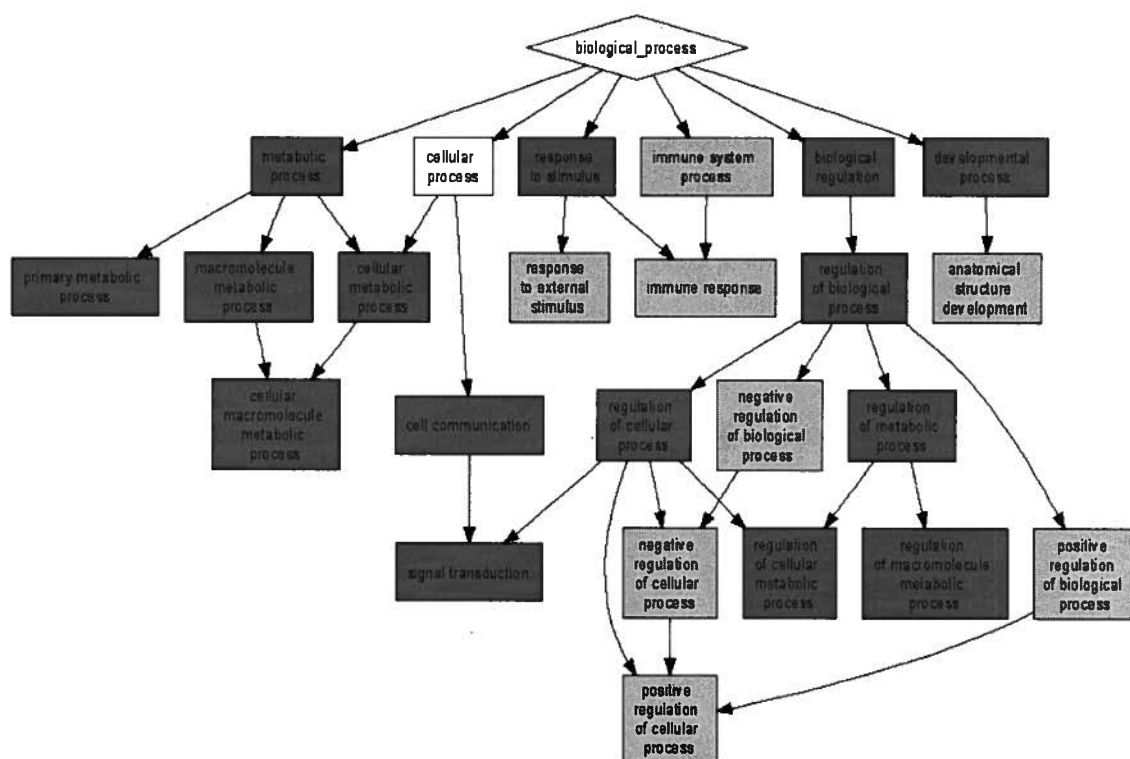


Figure 3.18: Gene Ontology highlighting the biological process significantly enriched in genes which are potentially regulated by STAT1 and Tel2.

3.6. Genome-wide analysis on function of genes potentially regulated by STAT1 and TF_x

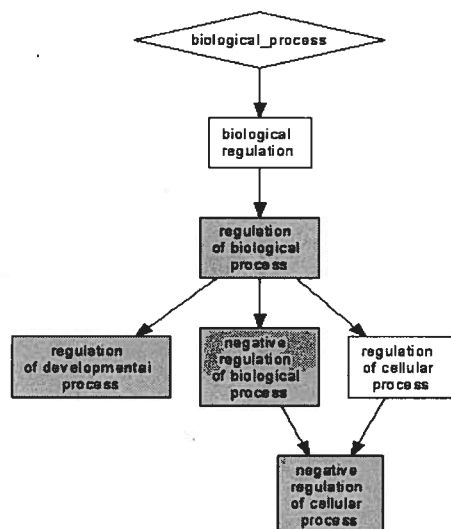


Figure 3.19: Gene Ontology highlighting the biological process significantly enriched in genes which are potentially regulated by STAT1 and AP1.

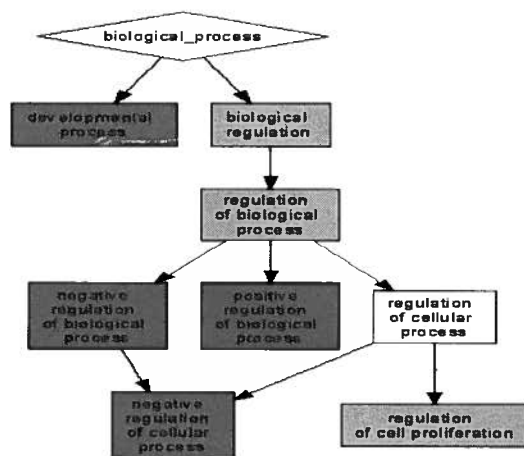


Figure 3.20: Gene Ontology highlighting the biological process significantly enriched in genes which are potentially regulated by STAT1.

3.7 The occurrence rate of binding sites corresponding to de novo discovered motifs

Here we checked whether the occurrence of binding sites corresponding to de novo discovered motifs are influenced by the location feature of the STAT1 sequences.

3.7.1 Categories of STAT1 sequences

For each STAT1 binding site detected by ChIP-Seq, we knew whether it is in a meaningful Me1 flanked region or whether it is close to a Pol2 binding site (i.e., have a Pol2 binding site within 2,500 bp). Therefore, we put the corresponding STAT1 sequences into four categories: Category 1, in Me1 flanked region and close to Pol2; Category 2, in Me1 flanked region and far from Pol2; Category 3, NOT in Me1 flanked region and close to Pol2; Category 4, NOT in Me1 flanked region and far from Pol2. Here, we used 9,992 STAT1 binding sites, and all the Pol2 binding sites detected by MACS and the 200~1,000bp Me1 flanked regions from all Me1 binding sites predicted by Robertson et al. (Robertson2008). The number of STAT1 binding sites belonging to each category can be visualized in the Venn diagram shown before (Figure 3.9).

3.7.2 Occurrence rate of binding sites corresponding to de novo motifs in STAT1 sequence of different categories

For each de novo predicted motif, we know whether it occurs in each one of the STAT1 sequence. Therefore, after putting the STAT1 sequences in different categories, we can check the occurrence rate of the binding sites corresponding to each motif in all the STAT1 sequences and in STAT1 sequences belonging to each of the four categories. Here, the occurrence rate of binding site corresponding to a motif x is defined as $\# \text{sequences checked which have binding site corresponding to motif } x / \text{total } \# \text{ sequences checked}$. Result is shown in Figure 3.21.

Figure 3.21 shows: 1) binding site corresponding to STAT1 motif has highest over all occurrence rate, followed by Nanog, HEB, Tel2 and AP1. 2) the occurrence rate of binding sites corresponding to different motifs are different. For example, binding site of STAT1 motif has highest occurrence rate (0.68) in the STAT1 sequences belonging to category "STAT1 not in flank and far from Pol2", binding site of AP1 motif has highest occurrence

3.7. The occurrence rate of binding sites corresponding to de novo discovered motifs

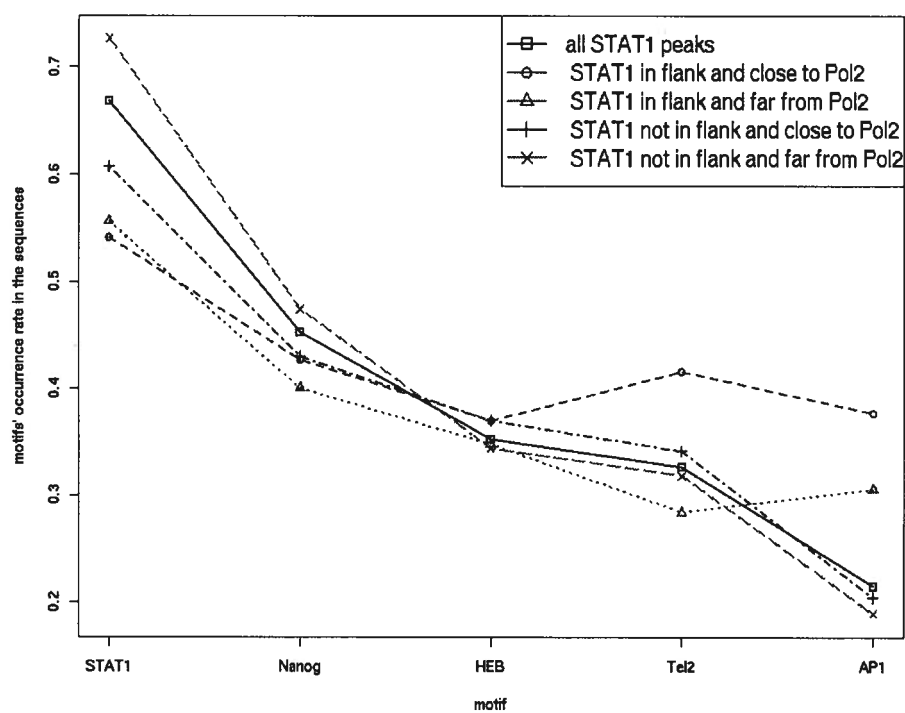


Figure 3.21: Occurrence rate of binding sites corresponding to de novo discovered motifs in all of the STAT1 sequences and in STAT1 sequences of different categories (categories are based on all predicted Pol2 binding sites and all predicted Me1 flanked regions).

3.7. *The occurrence rate of binding sites corresponding to de novo discovered motifs*

rate (0.38) in the STAT1 sequences belonging to category “STAT1 in flank and close to Pol2”.

Furthermore, for each motif, we did hypothesis test to check whether the occurrence rate of its corresponding binding site corresponding is significantly different in STAT1 sequences belonging to each pair of categories (we have 4 categories, so we did $\binom{4}{2}=6$ hypothesis tests for each motif_x). Basically, we assume that the number of occurrence of binding site corresponding to motif_x in a category follows a Binomial distribution with parameters n (total number of sequences in the category) and p (occurrence rate); the null hypothesis is that p of two categories of STAT1 sequences under test are the same. More details of this method are discussed in “Multiple test for proportions” of the Method chapter.

We found that at the significance level of 0.05, binding site of Tel2 has significantly different occurrence rate in STAT1 sequences belonging to the four different categories; binding site of STAT1 have significantly different occurrence rate in the five comparisons we did, except for the comparison between the sequence category “in Mel1 flanked region and close to Pol2 binding site” and the sequence category “in Mel1 flanked region and far from Pol2 binding site”. We conclude that the occurrence of binding sites corresponding to de novo discovered motifs are influenced by the location feature of the STAT1 sequences. Results are shown in Table 3.6.

Table 3.6: Comparison of each motif’s occurrence rate in STAT1 sequences of different categories.

Significant difference of binding site’s occurrence rate in two categories of STAT1 sequences	Motif				
	STAT1	Nanog	HEB	Tel2	AP1
InMelClosetoPol2 vs InMelFarfromPol2	NO	NO	NO	YES	YES
InMelflankClosetoPol2 vs NotinMelflankClosetoPol2	YES	NO	NO	YES	YES
InMelflankClosetoPol2 vs NotinMelflankFarfromPol2	YES	NO	NO	YES	YES
InMelflankFarfromPol2 vs NotinMelflankClosetoPol2	YES	NO	NO	YES	YES
InMelflankFarfromPol2 vs NotinMelflankFarfromPol2	YES	YES	NO	YES	YES
NotinMelflankClosetoPol2 vs NotinMelflankFafromPol2	YES	YES	NO	YES	NO

3.8. *The occurrence rate of binding sites corresponding to different combinations of de novo discovered motifs*

3.7.3 Occurrence rate of binding sites corresponding to de novo motifs in STAT1 sequence of different categories (categories based on Me1 flanked regions and Pol2 sites with more stringent criteria)

We selected top ~10,000 Pol2 binding sites according to the FDR for the Pol2 binding site predicted by MACS; and got top ~20,000 Me1 flanked regions for top Me1 binding sites with large number of short reads in them. We put STAT1 binding sites into four categories based on their location relationship with top Pol2 sites and top Me1 flanked regions. We checked the occurrence rate of binding sites corresponding to de novo motifs in STAT1 sequence of different categories again, in order to see whether the binding sites' occurrence rate change. Result is shown in Figure 3.22.

Comparing Figure 3.22 and Figure 3.21, we found: as we selected top Pol2 and top Me1 flanked regions, the criteria for the category STAT1 binding sites in Me1 flanked region and close to Pol2 became stringent, and the occurrence rate of the binding sites in this category changed. For example, the occurrence rate of binding site of STAT1 in this category dropped from 0.54 to 0.45; the occurrence rate of binding site of Tel2 increased from 0.45 to 0.55.

3.8 The occurrence rate of binding sites corresponding to different combinations of de novo discovered motifs

We have 5 motifs, so there are 32 (2^5) unique combinations of these motifs' occurrence or not within a sequence. We are interested to know, within each category, how the binding sites corresponding to different motif combinations occur.

We checked occurrence rate of binding sites corresponding to these combinations in all the sequences and in sequences belonging to different categories (occurrence rate of binding site corresponding to a motif combination = #sequence with binding site corresponding to motif combination in a category / total #sequences in a category). Result is shown in Figure 3.23.

Figure 3.23 shows that: 1) 15 out of the 32 combinations have highest occurrence rate in the category of "STAT1 in Me1 flanked region and close to Pol2" compared with other categories; and combination with all the motifs (combination 15) has the highest occurrence rate for this category comparing with other categories, showing that the DNA region within Me1

3.8. The occurrence rate of binding sites corresponding to different combinations of de novo discovered motifs

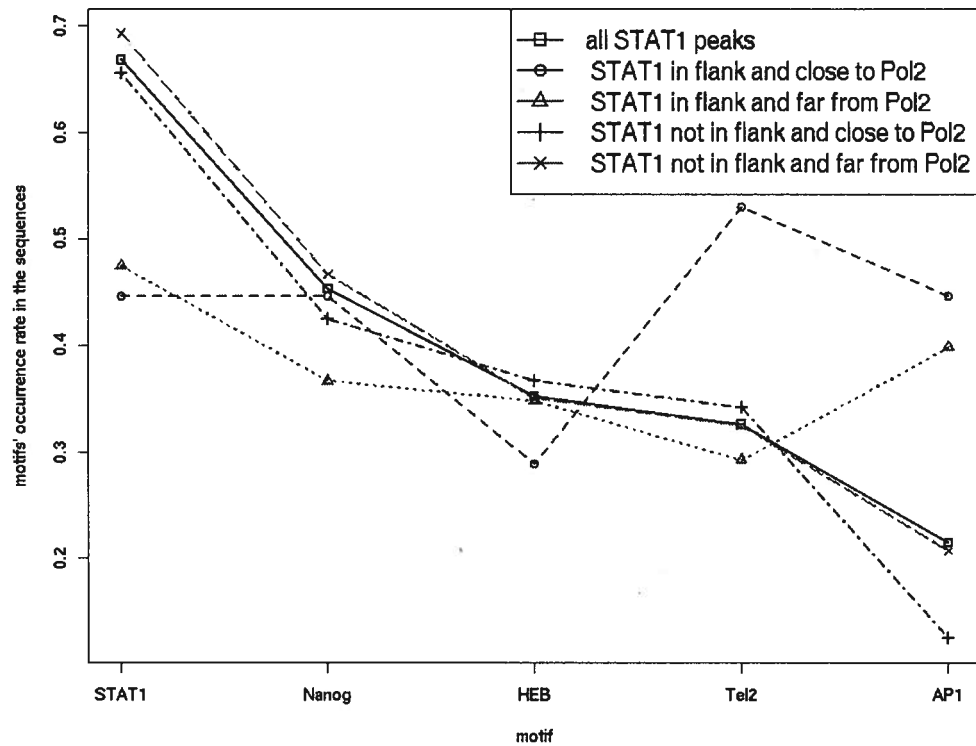


Figure 3.22: Occurrence rate of binding sites corresponding to de novo discovered motifs in all of the STAT1 sequences and in STAT1 sequences of different categories (categories are based on top predicted Pol2 binding sites and top predicted Me1 flanked regions).

3.8. The occurrence rate of binding sites corresponding to different combinations of de novo discovered motifs

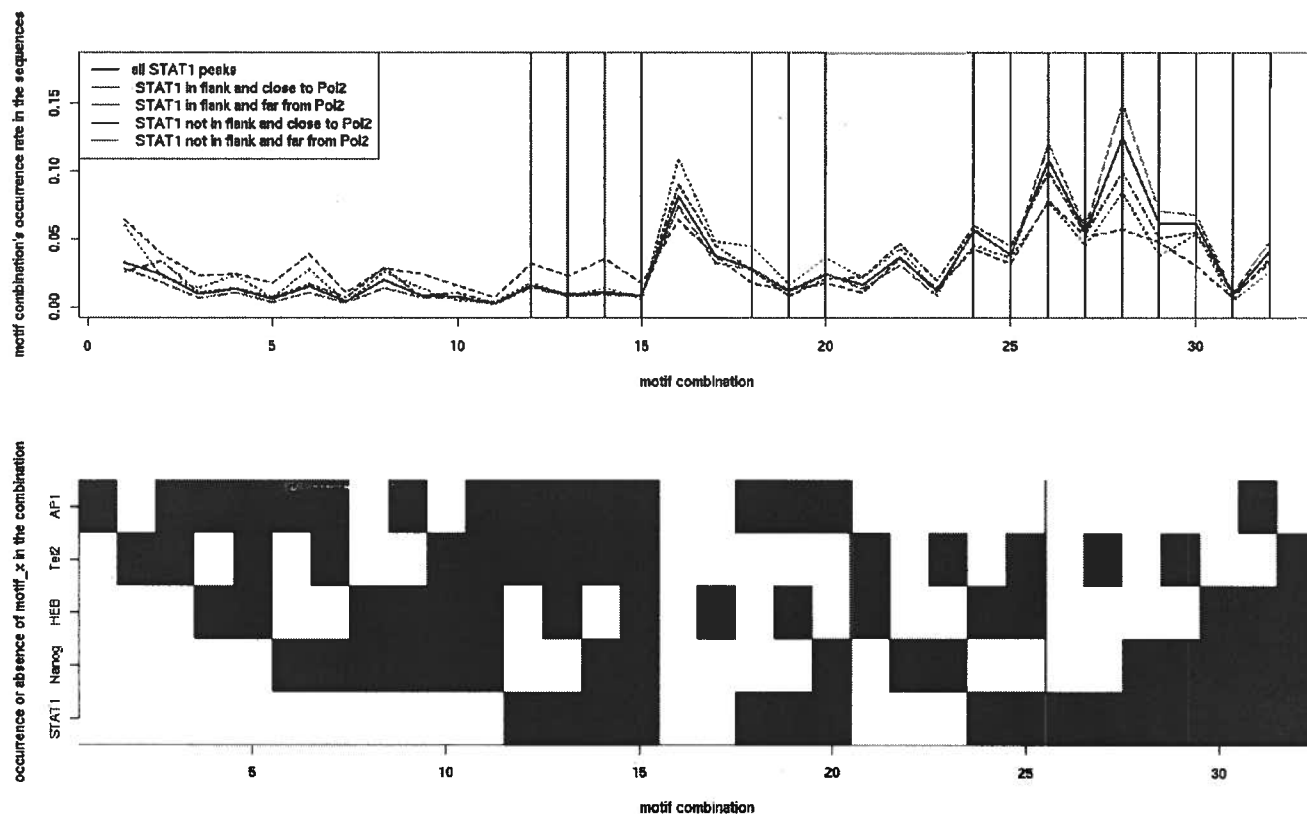


Figure 3.23: Top: occurrence rate of binding site corresponding to different motif combinations, the combinations whose occurrence rates are highest for the same category are put close to each other (here, categories are based on all Pol2 binding sites and all Me1 flanked regions). We use black line to indicate a motif combination if it has STAT1 occurring. Bottom: occurrence or absence of motif_x in the combinations. (Red color indicates the occurrence of a motif, white color indicates the absence of a motif; green lines are the border of the motif combinations: motif combinations whose occurrence rate are highest in the same category are put in the same chunk and are ordered as in the graph at top.)

flanked region and with Pol2 is where the binding sites are located most often. 2) Many combinations whose occurrence rate is highest in the category “STAT1 in Me1 flanked region and close to Pol2” have Tel2 and AP1, suggesting that Tel2 and AP1 may be important transcription factors. 3) Some combinations have highest occurrence rate in the category “STAT1 not in flank and far from Pol2”. We are not sure whether the predicted binding sites in this category really have biological function. 4) The combination with no motif in it has high occurrence rate in all the categories, indicating that our prediction on STAT1 binding site probably have noise, i.e. some of the locations predicted to be bound by STAT1 with ChIP-Seq are not actually bound by STAT1, and there is no binding site predicted for sequences around these locations.

3.9 Relating STAT1 binding sites with DE genes

Robertson et al. reported that the amount of STAT1 binding sites after IFN-gamma stimulation is about 4 times the amount of STAT1 binding sites before IFN-gamma stimulation (Robertson et al. [53]).

We know that TF regulates the transcription of a gene through binding to its promoter or enhancer region. We want to check in the IFN-gamma stimulated cell, whether the STAT1 binding is related with DE, i.e., whether the proportion of DE gene having STAT1 binding site in the promoter regions is higher than the proportion of all genes having STAT1 in the promoter regions.

3.9.1 DE genes detected on Chromosome 22 of IFN-gamma stimulated HeLa cell

In the work of Hartman et al., 63 genes showing differential expression (DE) after the IFN-gamma stimulation were identified on Chromosome 22 of HeLa cell. (Hartman et al. [24]).

3.9.2 DE genes detected for other three types of human cells under IFN-gamma stimulation

We also obtained time series microarray data for other three types of IFN-gamma stimulated human cells. These microarray data studied the gene expression on the whole genome scale. For each of three data set, we identified the DE genes as described in method section.

3.9. Relating STAT1 binding sites with DE genes

Table 3.7: Summary of number of total genes and number of DE genes in three microarray data sets

	Array_bloodcell	Array_skincell	Array_fibroblastcell
Total genes	11,680	8,759	14,319
DE genes	534	452	1,452

Table 3.7 summarizes the total number of genes studied in each microarray experiment (genes with too many missing value are not considered), and the number of DE genes detected from each microarray data set.

3.9.3 Intersection of DE genes in HeLa cell and DE genes in other three types of cells

The gene symbols for the genes studied with microarray experiment are known. We checked the intersections of gene symbols of DE genes in HeLa cell and in cells from the other three different tissues (peripheral blood mononuclear cell, skin cell and fibroblast cell).

Among the 63 DE genes detected for Chromosome 22 of HeLa cell, three show differential expression in the blood cell, none shows differential expression in the skin cell, six show differential expression in fibroblast cell. Here, the DE genes were only checked for Chromosome 22, therefore, we were not sure whether they can represent all the known genes in HeLa cell.

The number of DE genes detected for other three cell types are relatively large. We further checked the intersection of DE genes detected for the other three cell types. Only 6 genes are differentially expressed in all these types of cells, and none of the 6 genes are detected to be differentially expressed in HeLa cell. Figure 3.24 shows the intersection of the DE genes in the other three cells.

3.9.4 Proportion of DE gene promoters having STAT1 binding sites and proportion of all gene promoters having STAT1 binding sites

We checked whether the STAT1 binding detected by ChIP-Seq occurs more often in the promoters of DE genes than in promoters of all the genes.

We obtained 6,001bp promoter regions of DE genes detected on Chro-

Intersection of DE genes from cell of 3 tissues

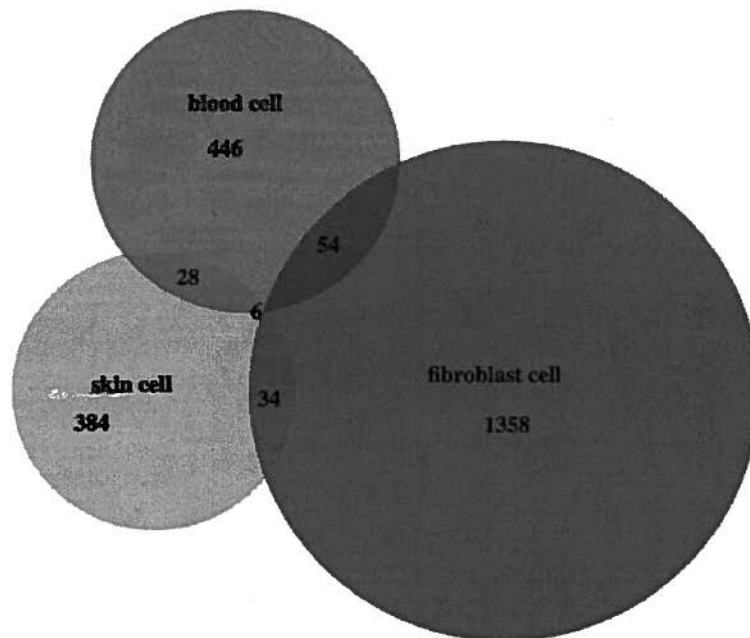


Figure 3.24: Intersection of DE genes found in three types of cells after IFN-gamma stimulation.

3.9. Relating STAT1 binding sites with DE genes

Table 3.8: Proportion of promoter of all genes with STAT1 binding sites and proportion of promoter of DE genes with STAT1 binding sites. (Note that a gene symbol may correspond to more than one predicted promoter, and we used all the non-redundant promoters)

	Promoters of all the genes	Promoters of DE genes
Number of promoters with STAT1 binding site	2,599	361
Total number of promoters	42,648	3,913
Proportion of promoters with STAT1 binding site	6.0%	9.2%

mosome 22 of HeLa cell and all the DE genes detected for other three types of cells (by merging the DE genes from three types of cells) and checked how many of them have STAT1 binding site; also, we obtained 6,001bp (4,000bp upstream and 2,000bp downstream) promoter regions of all the genes and counted how many of them have STAT1 binding site. Here, we used the 6,001bp region length because we found that as the downstream region length grows to 2,000 bp and the upstream region length grows to 4,000bp, the number of regions with STAT1 grows quickly (shown in Figure 3.6), and we wanted to use region which is long enough for finding most of the TF and at reasonable length at the same time.

None of the 63 DE genes detected on Chromosome 22 of HeLa cell have STAT1 binding in their promoter regions, while some DE genes detected for other three types of cells have STAT1 binding site in their promoter regions. Table 3.8 shows the proportion of DE genes with STAT1 binding in the promoter regions and proportion of all the human genes with STAT1 binding in the promoter regions. In the IFN-gamma stimulated cells, the proportion of DE promoter regions with STAT1 binding site is significantly higher than that of all promoter regions (we compared the two proportions with one tail z test at significance level of 0.05, as mentioned in the method section "Testing proportion of two samples").

There are only 6 genes that are differentially expressed in all three types of cells. We further checked whether there is STAT1 binding site occurring in the promoter region of these 6 genes (the promoter region we used is -400bp~+2,000bp of TSS). We found that 2 of these 6 genes have STAT1 binding in their promoter region, as shown in Table 3.9.

We acknowledge that no DE gene detected for Chromosome 22 in HeLa cell has STAT1 binding in the promoter region. However, through the analysis of STAT1 binding and the DE genes in cells other than HeLa, we infer

3.9. Relating STAT1 binding sites with DE genes

Table 3.9: Genes differentially expressed in all 3 types of cells and whether there is STAT1 binding site in their promoter regions

gene symbol of DE gene	with STAT1 binding site in the promoter region
OAS2	NO
IFRD1	NO
INDO	YES
INHBA	YES
MGLL	NO
DUSP6	NO

that STAT1 binding is related with gene's differential expression. Our analysis on DE genes and STAT1 binding is limited here, because many DE genes in the HeLa cell may be different from the DE genes we detected for the other three types of cells.

We hope that gene expression experiment can be done for all the known genes in IFN-gamma stimulated HeLa cell in the future, so that the STAT1 ChIP-Seq data and the gene expression data are more comparable.

Chapter 4

Discussion

4.1 The binding sites of TFs

The ChIP-Seq experiment enables us to find binding sites of a specific TF (STAT1 in our study) on the whole genome scale. Based on the 401bp sequences around STAT1 binding sites, we predicted several de novo motifs which may be corresponding to TFs collaborating with STAT1 in gene regulation.

4.1.1 The overlap of binding site locations of de novo predicted motifs

It has been verified in the wet lab experiment that the binding sites of two TFs can overlap with each other. In our analysis, we found that there are overlaps between locations of binding sites corresponding to STAT1 GAS motif and other motif_x. For example, the predicted binding site of STAT1 and Nanog overlap in almost 2,000 locations. We are not sure whether TFs bind to the overlapping sites as often as predicted.

4.1.2 What are the non-coding regions in the genome? Are TF binding in regions far from genes regulating the gene transcription?

The haploid human genome has just over 3 billion DNA base pairs. In the genome, there are about 25,000 genes whose medium length is 20,000 base pairs. The genes occupy only about 1/6 of the genome. Moreover, there are introns and other untranslated regions in the genes.

Through analysis, we found that many STAT1, Pol2 binding sites and Mel flanked regions are distal from TSS, i.e., they are not located in any gene regions. We want to know what fraction of the TF binding in the non-coding regions of the genome are functional in regulating the genes' expression, and what fraction of them are "junk" in view of regulating gene

transcription. More expression experiments need to be done to answer the question.

4.2 Uncertainty in the specificity of TF binding

GO analysis of the genes potentially regulated by STAT1 and TF_x, as well as literature review gave us evidence that the binding sites corresponding to motif_x found in the STAT1 sequences form cis-regulatory module with STAT binding sites. However, it is not clear to us which TFs actually correspond to the motif_x.

4.2.1 Uncertainty in deciding which TF has binding site similar to the de novo predicted motif

We are not certain about what exact TF has the binding site similar to the de novo predicted motif: transcription factors with similar binding domain are put into the same family, and some TFs in the same family have similar DNA binding sites/motifs. For example, the motifs of STAT1 and STAT5 from STAT family are very similar, the motifs of AP1 and Bach2 from bZIP family are very similar (refer to Table 3.3).

In our analysis, we call the motif_x using the name of TF whose binding site motif is most similar to motif_x. But it is possible that several TFs whose binding site motifs are similar to motif_x cooperate with STAT1 in regulating different genes.

4.2.2 Uncertainty in the specificity of how TFs collaborate

It is not clear to us, whether the TFs collaborate in a very precise way (e.g., transcription factor A collaborate with transcription factor B, and they regulate a gene's expression by binding to its promoter region together), or whether a TF can sometimes collaborate with several TFs from another family in regulating a gene's transcription (e.g., transcription factor A can collaborate with several transcription factors from the same family of transcription factor B in regulating gene transcription).

ChIP-PCR can help to answer our question for a specific gene and ChIP-Seq can help to answer our question on the whole genome scale.

4.3 The condition of TF binding

We know there are many factors influencing the TF binding to DNA. For instance, STAT1 can bind to DNA after it is phosphorylated; binding of a TF may depend on the availability of its cofactors; promoter and enhancer of genes actively transcribed are marked by histone methylation. Therefore, given the genome sequence and the predicted binding sites of a TF, it is still difficult to tell whether the TF binds to the predicted binding sites. Further studies on the mechanism of TF binding need to be conducted.

4.4 Potential wet-lab experiment

4.4.1 Potential ChIP-Seq experiment

By de novo motif prediction on 9,992 401bp STAT1 sequences, we obtained 4 motifs, whose corresponding TFs potentially cooperate with STAT1 in gene regulation. The TFs that have motifs most similar to the de novo predicted motifs are: AP1, Nanog, HEB and Tel2. It is probable that the binding sites of motif_x constitute cis regulatory module with binding sites of STAT1, and that TF_x collaborate with STAT1 in regulating gene transcription.

We looked for ChIP-Seq or ChIP-chip experiment for the TFs corresponding to these motifs in GEO database. We only found ChIP-chip data for HEB. Analyzing HEB data and STAT1 data, we found 21 regions on Chromosome 19 that have HEB and STAT1 binding sites close to each other.

Result through literature review showed that Nanog, TFs from Ets family, AP1 collaborate with STAT1 in regulating gene expression, which verified our prediction to some extent. Yet, searching for papers is a time-consuming process, and usually the regulation of only one gene is discussed in a paper. Therefore, the literature review can not provide us enough information of how two TFs collaborate with each other on a whole genome scale.

Therefore, we found it necessary and helpful to carry out ChIP-Seq analysis for the TFs corresponding to de novo predicted motif_x (TF_x). With more ChIP-Seq experiment on the TFs, we can get to know the binding sites of TF_x and check: 1) whether there is real transcription factor binding for the predicted binding site of motif_x; 2) how the spacial relationship of TF_x and STAT1's binding sites is, and how often their binding site locations overlap.

Note that through analyzing the microarray data, we found the DE genes in different cell types under IFN-gamma stimulation are different. Therefore,

the binding of transcription factors in the promoter regions which leads to DE may be different for different types of cell. In order to make the new experiment consistent with STAT1 ChIP-Seq experiment, we suggest that ChIP-Seq experiment be carried out for the IFN-gamma induced HeLa cell.

4.4.2 Potential gene expression experiment

The ChIP-Seq experiment for STAT1 was done for the HeLa cell under IFN-gamma stimulation. In our analysis, we have used gene expression data of HeLa cell under IFN-gamma stimulation, but it was only for genes on Chromosome 22. Besides that data set, we have also used gene expression data of other types of IFN-gamma stimulated human cells.

Our analysis showed that no DE gene detected for Chromosome 22 in HeLa cell has STAT1 binding in the promoter region. However, some DE genes detected for other three types of human cells have STAT1 binding site in their promoter regions and the proportion of DE gene promoter regions with STAT1 binding site is higher than that of all promoter regions. We inferred that STAT1 binding may be related with gene's differential expression.

Our analysis also showed that the intersection of DE genes occurring in different cell types after IFN-gamma stimulation is small. Therefore, we hope that time-course microarray experiment for all the genes in IFN-gamma stimulated HeLa cell will be available in the future. By studying the microarray data of IFN-gamma stimulated HeLa cell and ChIP-Seq data of STAT1 binding in IFN-gamma stimulated HeLa cell together, we can: 1) check the proportion of the differentially expressed genes having STAT1 binding in the promoter region; 2) look into the gene coding region and possible enhancer region of the differentially expressed genes and check whether there is STAT1 binding; 3) compare the frequency of STAT1 binding in up-regulated and down-regulated genes; 4) identify whether there is cis-module in the promoter region of genes with similar expression pattern; 5) compare the DE genes in HeLa cell and the DE gene in other IFN-gamma stimulated cell, and identify which genes always show differential expression regardless of the cell type.

4.5 Use of R, Perl and SQL

In this thesis, we used R, Perl and MySQL to do data analysis. All the software packages are available for free.

4.5. *Use of R, Perl and SQL*

We used R and packages based on R to perform statistical analysis and visualization for large scale data; we used Perl to extract useful contents from the program output or from data tables; database is useful and efficient for handling data tables, such as storing, extracting and combining information of different tables.

Bibliography

- [1] Ashburner M. et al Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).
- [2] Babu M.M. et al. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* **14**, 283-291 (2004).
- [3] Bailey T.L. and Elkan Charles Fitting an mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol* (1994)
- [4] Bailey T.L. et al. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**, W369-W373 (2006).
- [5] Bauer S. et al. Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**, 1650-1651 (2008).
- [6] Berman P. et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *PNAS* **105**, 757-766 (2008).
- [7] Blackwood E.M. and Kadonaga J.T. Going the Distance: A Current View of Enhancer Action. *Science* **280**, 60-63 (1998).
- [8] Bluthgen N. et al. Biological profiling of gene groups utilizing Gene Ontology. *Genome Informatics* **16**, 106-115 (2005).
- [9] Brooker R.J. Genetics: analysis and principles (2nd edition). *McGraw-Hill* (2005).
- [10] Bulyk M.L. et al. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research* **30**, 1255-1261 (2002).
- [11] Chen W. et al. Saturation mutagenesis of a yeast *his3* "TATA element": Genetic evidence for a specific TATA-binding protein. *Proc. Natl. Acad. Sci.* **85**, 2691-2695 (1988).

- [12] Crooks G.E. et al. WebLogo: A Sequence Logo Generator. *Genome Research* 14, 1118-1190 (2004).
- [13] Davidson E.H. Genomic regulatory systems: development and evolution. (2001).
- [14] Duff J.L. et al. Pervanadate mimics IFN γ -mediated induction of ICAM-1 Expression via activation of STAT1 proteins. *J. Invest. Dermatol.* **108**, 295-301 (1997).
- [15] Eden E. et al. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
- [16] Fejes A.P. et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729-1730 (2008).
- [17] Fujio Y. et al. Signals through gp130 upregulate *bcl-x* gene expression via STAT1-binding cis-element in cardiac myocytes. *J. Clin. Invest.* **99**, 2898-2905 (1997).
- [18] Gardini A. et al. AML1/ETO oncoprotein is directed to AML1 binding regions and co-localizes with AML1 and HEB on its targets. *PLoS Genetics* **4**, 1-7 (2008).
- [19] Ganster R.W. et al. Complex regulation of human inducible nitric oxide synthase gene transcription by Stat 1 and NF- κ B. *PNAS* **98**, 8638-8643 (2001).
- [20] Gao J. et al. An interferon-gamma-activated Site (GAS) is necessary for full expression of the mouse iNOS gene in response to IFN-gamma and lipopolysaccharide. *J. Biol. Chem.* **272**, 1226-1230 (1997).
- [21] Gariglio P. et al. Clustering of RNA polymerase B molecules in the 5' moiety of the adult β globin gene of hen erythrocytes. *Nucleic Acids Research* **9**, 2589-2598 (1981).
- [22] Gorodkin J. et al. Displaying the information contents of structural RNA alignments: the structure logos. *CABIOS* **13**, 583-586 (1997).
- [23] Grillot D.A. et al. Genomic organization, promoter region analysis, and chromosome localization of the mouse *bcl-x* gene. *The Journal of Immunology* **158**, 4750-4757 (1997).

- [24] Hartman S.E. et al. Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes and Dev.* **19**, 2953-2968 (2005).
- [25] Heintzman N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**, 311-318 (2007).
- [26] Hsu F. et al. The UCSC known genes. *Bioinformatics* **22**, 1036-1046 (2006).
- [27] Ji H. et al. An integrated software system for analyzing ChIP-chip and ChIP-Seq data. *Nature Biotechnology* **26**, 1293-1300 (2008).
- [28] Kanda N. et al. Histamine enhances the production of human *beta*-defensin-2 in human keratinocytes. *Am J Physiol Cell Physiol* **293**, C1916-C1923 (2007).
- [29] Kang Y. et al. A self-enabling TGF response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells. *Molecular Cell* **11**, 915-926 (2003).
- [30] Kent W.J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
- [31] Kielbasa S.M. et al. Genome-wide analysis of functions regulated by sets of transcription factors. *German Conference on Bioinformatics 2004* **54**, 105-113 (2004).
- [32] Koch F. et al. Genome-wide RNA polymerase II: not genes only! *Cell* **33**, 265-273 (2008).
- [33] Kornberg, R.D. The molecular basis of eukaryotic transcription. *PNAS* **32**, 12955-12961 (2007).
- [34] Kristof A.S. et al. Mitogen-activated protein kinases mediate activator protein-1-dependent human inducible nitric-oxide synthase promoter activation. *J. Biol. Chem.* **276**, 8445-8452 (2001).
- [35] Kutach A.K. et al. The downstream promoter element DPE appears to be as widely used as the TATA Box in Drosophila core promoter. *Molecular and Cellular Biology* **20**, 4754-4764 (2000).
- [36] Latchman D.S. Transcription Factors: an Overview. *Int. J. Biochem. Cell Biol.* **29**, 1305-1312 (1997).

- [37] Latchman D.S. Inhibitory Transcription Factors. *Int. J. Biochem. Cell Biol.* **28**, 965-974 (1996).
- [38] Launoit Y.D. et al. The transcription of the intercellular adhesion molecule-1 is regulated by Ets transcription factors. *Oncogene* **16**, 2065-2073 (1998).
- [39] Li L. GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *Journal of Computational Biology* **16**, 317-329 (2009).
- [40] Louie M.C, Yang H.Q. et al. Androgen-induced recruitment of RNA polymerase II. *PNAS* **100**, 2226-2230 (2003).
- [41] Lowenstein et al. Macrophage nitric oxide synthase gene: Two upstream regions mediate induction by interferon γ and lipopolysaccharide. *Biochemistry* **90**, 9730-9734 (1993).
- [42] Luger K. et al. Crystal structure of the nucleosome core particle at 2.8Å resolution. *Nature* **389**, 251-260 (1997).
- [43] Mahony S. and Benos P.V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research* **35**, W253-W258 (2007).
- [44] Mardis E.R. ChIP-Seq: welcome to the new frontier. *Nature methods* **4**, 613-614 (2007).
- [45] Messina D. et al. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.* **14**, 2041-2047 (2004).
- [46] Mineshiba J. et al. Transcriptional regulation of *beta*-defensin-2 by lipopolysaccharide in cultured human cervical carcinoma (HeLa) cells. *FEMS Immunology and Medical Microbiology* **45**, 37-44 (2005).
- [47] Ohler U. et al. Computational analysis of core promoters in the Drosophila genome. *Genome Biology* **3**, 0087.1-0087.12 (2002).
- [48] Pabo C.O. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053-1095 (1992).
- [49] Dhaeseleer P. What are DNA sequence motifs? *Computational Biology* **24**, 423-425 (2006).

- [50] Pennisi, E. A Low Number Wins the GeneSweep Pool *Science* **300**, 1484 (2003).
- [51] Radonjic M. et al. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Molecular Cell* **18**, 171-183 (2005).
- [52] Ramana C.V. et al. Complex roles of Stat1 in regulating gene expression. *Oncogene* **19**, 2619-2627 (2000).
- [53] Robertson A.G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651-657 (2007).
- [54] Robertson A.G. et al. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Research* **18**, 1906-1907 (2008).
- [55] Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *TIBS* **21**, 327-334 (1996).
- [56] Roth F.P. et al. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939-945 (1998).
- [57] Rozowsky J. et al. PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. *Nature Biotechnology* **1**, 66-75 (2009).
- [58] Sandelin A. et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **32**, D91-D94 (2004).
- [59] Schroder K. et al. Interferon-gamma: an overview of signals, mechanisms and functions. *Journal of Leukocyte Biology* **75**, 163-189 (2004).
- [60] Simes R.J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754 (1986).
- [61] Simonoff J.S. Analyzing categorical data. (2003).
- [62] Smith A.D. et al. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *PNAS* **103**, 6275-6280 (2006).

- [63] Singer V. et al. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcription activation. *GENES & DEVELOPMENT* **4**, 636-645 (1990).
- [64] Staden R. Methods for calculating the probabilities of finding patterns in sequences. *CABIOS* **5**, 89-96 (1989).
- [65] Sun Y. et al. Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation. *PLoS ONE* **3**, e3406 (2007).
- [66] Symes A. et al. STAT proteins participate in the regulation of the vasoactive intestinal peptide gene by the ciliary neurotrophic factor family of cytokines. *Mol. Endocrinol.* **8**, 1750-1763 (1994).
- [67] Symes A. et al. Integration of Jak-Stat and AP-1 signaling pathways at the vasoactive intestinal peptide cytokine response element regulates ciliary neurotrophic factor-dependent transcription. *J. Biol. Chem.* **272**, 9648-9654 (1997).
- [68] Thijs G. et al. INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics* **18**, 331-332 (2002).
- [69] Thompson W. et al. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research* **31**, 3580-3585 (2003).
- [70] Valouev A. et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5**, 829-835 (2008).
- [71] Wingender E. et al. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research* **24**, 238-241 (1996).
- [72] Zeitlinger et al. RNA Polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genetics* **39**, 1512-1516 (2007).
- [73] Zhang X. et al. PICS: Probabilistic inference for ChIP-seq. arXiv:0903.3206v1 [q-bio.GN] (2009).
- [74] Zhang Y. et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

Appendix A

Parameter setting for running MACS

The parameter setting for running MACS is list below:

```
#default setting of mfold will give warning Fewer paired peaks (250)
than 1000! Model may not be build well! Lower your MFOLD parameter
may erase this warning. So I will use 250 pairs to build model!
macs -t ../../rawdata/STAT1_stimulated/STAT1_stimulated.bed -c ../../raw-
data/STAT1_input/STAT1_input.bed -name=mfold_25 -mfold=25 -tsize=28
macs -t ../../rawdata/Pol2_stimulated/Pol2_stimulated.bed -c ../../raw-
data/Pol2_input/Pol2_input.bed -tsize=28
```


Appendix B

Parameter setting for running DecoyMasker

The parameter setting for running DecoyMasker is list below:

```
CREADBINDIR=/export/home/kaida/cread/cread-0.84/bin
FADIR=/export/home/kaida/seqs
FA=length400all
# Three-stage decoymasking
R=2
W=7
DMFA2=$FA.1st_stage_decoymasker_w$W_r$R.fa
echo "Decoymasker, w=7, r=2 >"
echo $CREADBINDIR/decoymasker $FADIR/$FA -w $W -r $R -l $FADIR/
$FA.dm_log_r2 -o $FADIR/$DMFA2
$CREADBINDIR/decoymasker $FADIR/$FA -w $W -r $R -o $FADIR/$DMFA2
R=3
W=10
DMFA3=$FA.2_stage_decoymasker_w$W_r$R.fa
echo "Decoymasker, w=15, r=3 >"
echo $CREADBINDIR/decoymasker $FADIR/$DMFA2 -w $W -r $R -l
$FADIR/$FA.dm_log_r3 -o $FADIR/$DMFA3
$CREADBINDIR/decoymasker $FADIR/$DMFA2 -w $W -r $R -o $FADIR/$DMFA3
R=4
W=13
DMFA4=$FA.3rd_stage_dmasker_w$W_r$R.fa
echo "Decoymasker, w=15, r=3 >"
echo $CREADBINDIR/decoymasker $FADIR/$DMFA3
-w $W -r $R -l $FADIR/$FA.dm_log_r4 -o $FADIR/$DMFA4
$CREADBINDIR/decoymasker $FADIR/$DMFA3 -w $W -r $R -o $FADIR/$DMFA4
```

Appendix C

Parameter setting for running GADEM

The parameter setting for running GADEM is list below:

```
MAXGAP=2
EM=40
FRACEM=0.3
PV=0.0002
INPUT=refined400all.fa
REPORT=$INPUT.fracEM$FRACEM.minN$MINN.pv$PV.maxgap
$MAXGAP.de_novo
nice -n 19 ../../kaida_software/bin/gadem -fseq $INPUT -fout $REPORT
-em $EM -fracEM $FRACEM -pv $PV -maxgap $MAXGAP -verbose 1 -
minN 200
```