Transcription factor binding motif analyses in two biological systems

by

Shilu Zhang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

May, 2014

Abstract

The mechanism of gene regulation is a crucial problem in current computational biology. Chromatin-immunoprecipitation microarray (ChIP-chip) is a technique used to study transcriptional regulation by identifying the binding regions of specific transcription factors. In this thesis, we focus on the binding of transcription factors to upstream region motifs to understand the mechanism of gene regulation.

Sonic hedgehog (Shh) signals direct digit number and identity in the vertebrate limb via Gli transcription factors. We sought to identify key Gli binding motifs in Gli binding regions through whole-genome ChIP-chip in the developing mouse limb. Through de novo motif discovery method, we found that there were specific DNA motifs enriched in different expression domains of the developing limb. A novel motif in Gli binding regions is highly likely to be a functional element. In addition, we noted that there was no statistically significant difference of quality of Gli motifs in Gli binding regions associated with genes expressed in different domains. The quality of Gli motif might not be a factor that influences the expression of genes in different domains.

ABSTRACT

Myc transcription factor, produced by the MYC oncogene, has the ability to activate and repress gene transcription. Elevated expression of Myc transcription factor is frequently found in cancers. We conducted antibody-specific motif analysis with application to human MYC transcription factor by using high-throughput genomic approaches. Chromatin immunoprecipitation was performed using two different anti-Myc antibodies in human P493-6 B cells, Santa Cruz (SC) antibody and Epitomics (Epit) antibody. Intersection of the two peak lists from SC antibody and Epit antibody identified 885 common Myc binding regions in both data sets. With the average probe intensities of the ChIP samples, we found no statistically significant difference between the binding intensity of probe with ChIP sequences immunoprecipitated by SC antibody and Epit antibody. There was linear increasing relationship between Epit antibody and SC antibody. Furthermore, we identified that Sfpi1 motif might be specifically bound to Myc-SC antibody and involved in differentiation or activation of B-cells.

Primary Reader: Dr. Hongkai Ji

Secondary Reader: Dr. Jiang Qian

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Hongkai Ji for support of my work. His guidance helped me in all the time of research and writing of this thesis. Without his supervision and constant help, this thesis would not have been possible. He has been a great mentor for me. With one year's work, I have learned a lot from him and developed great interest in genomics. He also gave me precious advice on my future career. Thank you very much for encouraging my research and for allowing me to grow into a graduate.

I would like to thank my thesis reader Dr. Jiang Qian. Thank you very much for spending time reading my thesis and offering me advice. Your suggestions are very priceless for me.

I would like to thank Dr.Brian Caffo and Dr. Karen Bandeen-Roche for providing me such an opportunity to study in the Department of Biostatistics at Johns Hopkins Bloomberg School of Public Health. I have learned a lot and practiced myself during the two years' study. Many thanks to all the faculty and students in the department. Thank you very much for your help and I really enjoy my life here. In particular, I

ACKNOWLEDGMENTS

am grateful to Yingying Wei for teaching me so much knowledge about data analysis and enlightening me the first glance of research.

Last but not the least, I would like to thank my family for supporting me through my life. I will be grateful forever for your love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The mechanism of gene regulation is a crucial problem in current computational biology. ChIP-chip and ChIP-seq are powerful technologies to study transcriptional regulation in complex genomes. ChIP-chip experiments consist of chromatin immunoprecipitation (ChIP) of transcription factor-bound genomic DNA followed by high density oligonucleotide hybridization (chip) of the IP-enriched DNA [1]. ChIP-Chip technology can identify transcription factor binding sites on a genome-wide basis. ChIP-sequencing (ChIP-seq), is a new approach used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. Using these technologies, studies of protein-DNA interactions provide information on cis-regulatory circuitry underlying various cellular processes [2].

1.1 ChIP-chip

The chromatin immunoprecipitation on microarray (Chip-chip) is used to investigate interactions between DNA and proteins in vivo. The technology of ChIP-chip enables large-scale screening for the binding regions of specific transcription factors at a resolution of $\sim 1 - 2$ kb [3]. In a ChIP experiment for DNA-binding proteins, a transcription factor of interest is crosslinked to DNA by a formaldehyde fixation [4]. The chromatin is sheared by sonication into small fragments of average length 1kb, and the fragments bound by the transcription factor are precipitated using specific antibodies for the transcription factor. In the next “chip step”, the crosslinks of DNA-protein complexes are reversed and the DNA strands are released [5]. After an amplification and denaturation step, IP-enriched DNA is fluorescently labeled and hybridized to chips containing single-stranded DNA sequences [1]. As a result, the bound probes are expected to occur in small clusters, which we refer to as peaks [1]. By comparison between the ChIP sample and control sample, one can identify which parts of the genome are bound by the transcription factor. Therefore, ChIP-chip experiments can provide valuable information for locating cis-regulatory elements in the genome [3].

CisGenome is an integrated software system for analyzing ChIP-chip data. To analyze ChIP-chip data in CisGenome, one usually goes through data exploration, quantile normalization, binding regions detection, adding gene annotations and finding enriched sequence motifs [2]. Binding regions are detected using a moving average

(MA) method [6]. Studies for this thesis are mainly based on the ChIP-chip analysis [2].

1.2 ChIP-seq

In ChIP-seq, the DNA fragments of interest are sequenced directly instead of being hybridized on an array. Compared to ChIP-chip, ChIP-seq technology can provide higher resolution, less noise and greater coverage for identification of protein-DNA interactions [2,4].

In a ChIP experiment, DNA fragments associated with a specific protein are enriched. An antibody specific to the protein of interest is used to immunoprecipitate the DNA-protein complex. Finally, the crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein [4]. Oligonucleotide adaptors are added to the released DNA strands to enable massively parallel sequencing. After size selection, all the resulting ChIP-DNA fragments are sequenced simultaneously using a genome sequencer. There are many sequencing methods, such as Illumina Genome Analyzer, Applied biosystems' SOliD and the Helicos platform [4]. The Genome Analyzer and SOliD platforms generate 100-400 million reads in a single run with an accuracy rate up to 98% [4]. With higher resolution, fewer artefacts and a larger dynamic range, ChIP-seq technology therefore provides substantially improved data [4].

CisGenome can analyze ChIP-seq data as well. Analysis of a ChIP-chip experiment begins with aligning reads to the genome and finding binding regions. The predicted regions can then be used for downstream analyses, including motif discovery and annotation retrieval.

1.3 Motif analysis

Transcription factors are proteins required to initiate or regulate transcription in eukaryotic cells [3]. Transcription factors bind to specific DNA patterns in the transcription regulatory region (TRR) of genes and either induce or repress the transcription of these genes by recruiting other proteins [7]. The loci on the DNA sequences that transcription factors bind to are called cis-regulatory elements [3]. The binding sites of the same transcription factor show a significant sequence conservation, which is called a transcription factor binding motif (TFBM) or binding consensus with length of 5-20 bases long [7]. Cis-regulatory modules (CRM) are DNA sequences where a number of transcription factors can bind and regulate expression of nearby genes, i.e., combinatorial patterns of DNA motifs. Different transcription factors may have different binding motifs. A motif can be represented either by a consensus sequence, an alignment matrix, a frequency matrix, or a weight matrix [3]. Sequence logos can be used to visualize the motifs. Multiple transcription factors can bind cooperatively to a cis-regulatory element, which contains several different binding motifs that are closely clustered together [3]. Transcription regulation not only relies on the combination of the TFs involved, but also on the number of site copies in the upstream regions [7]. Characterizing the motifs of transcription factors and searching for locations of transcription factor binding sites are of great importance for understanding gene regulation mechanisms in response to developmental changes [7].

Motif discovery is an approach to find both the motif patterns and the locations of transcription factor binding sites (TFBSs) in DNA sequences [3]. De novo method is a way of discovering unknown motifs and their corresponding TFBSs that are "enriched" in a set of upstream sequences [7]. Compared to random sequences, the sites bound by the same transcription factor are enriched in the set of genes that are coregulated by this TF [3]. When a group of coregulated genes is available, we could look for their common regulatory mechanisms [3]. If coregulation of these genes is resulted from the binding of a common set of transcription factors, TFBSs should be found enriched in the surrounding DNA sequences of these coregulated genes [3]. By looking for overrepresented sequence patterns in the genomic regions near these genes, we can infer both transcription factor binding motifs and their positions [3]. If the binding motif of a transcription factor is known from biological experiments, we can score sequence patterns of the motif and hence predict TFBSs by scanning genomic DNA sequences [3]. This method is called known motif mapping. Motif discovery is of great importance since the binding of transcription factors to upstream region motifs is crucial to understanding the mechanism of gene regulation [3].

Chapter 2

DNA motif analysis in Gli binding regions

2.1 Introduction

In the developing limb, Sonic Hedgehog (Shh) signaling acts as a morphogen to specify the final digit pattern and a growth factor by directing a complex transcription response [8]. Sonic hedgehog (Shh), secreted by a discrete posterior organizing center, the zone of polarizing activity (ZPA), directs digit number and identity in the vertebrate limb [9]. Shh signaling is mediated via Gli transcription factor, which contains zinc finger domain, through a Gli-consensus binding sequence, TGGGTGGTC [8,9]. However, the mechanisms underlying the transcriptional responses are poorly understood. Vokes et al. have predicted 205 putative limb target genes through intersection of DNA binding data with gene expression profiles [9]. The binding regions that correspond to the 205 unique genes represent the core set of candidates under direct Gli transcriptional control in the developing limb [9]. The average size of Gli binding regions by ChIP-on-chip ranges from 200 to 3000 bp. For a further study on the Gli target genes, our collaborators have identified the gene expression patterns for the 205 candidate Shh target genes. They have clustered gene expression patterns into 4 expression categories, genes expressed in the posterior and the posterior-distal limb, genes expressed in the posterior-proximal limb, genes expressed in the central portion of the limb, genes expressed with multiple spatially distinct domains where one or more expression domain is within the overall Sonic Hedgehog-responsive region in the mouse limb. Interestingly, they have noted that the different groups have distinct requirements for Shh signaling. Shh signaling is required to maintain expression for genes within the posterior-distal group. Whereas, most genes within the other domains require Shh signaling only for initiating gene expression.

In this thesis, our aim is to check if there are any specific DNA motifs enriched in individual gene expression categories and the quality of Gli motifs in Gli binding regions. Different DNA motifs might be associated with different expression groups. Gli transcription factor, associated with other proteins, might bind to different DNA motifs, activating transcription of genes in different domains. The quality of Gli motifs in Gli binding regions might be statistically different between groups of genes expressed in the Shh-responsive region and genes that are not expressed in the Shh-responsive region. Would this difference have an influence in the expression of genes in different domains?

2.2 Methods

2.2.1 Data generation

We downloaded the list of 656 peak-gene pairs associated with 261 unique genes in the limb and list of 396 peak-gene pairs that contained Gli motifs associated with 205 unique genes in the limb from supplemental material in Vokes et al., 2008. The 656 binding regions were the total binding regions associated with differentially expressed limb genes, of which 396 binding regions containing Gli motifs. The original ChIP-chip datasets were generated using Affymetrix tilling arrays [10]. Raw data were quantile normalized and binding regions were determined using the new version of TileMap incorporated into CisGenome using a moving average (MA) method [6]. Details of generation of the two peak lists were provided in supplemental materials in Vokes et al., 2008.

Our collaborator Dr. Vokes provided us with 10 groups of genes, which were identified in Vokes et al., 2008 (Table 2.1). Included in each group are the gene names and RefSeq ID (as defined by Vokes et al., 2008). Group 1 consists of 58 genes predominately expressed in the Sonic Hedgehog-responsive region in the mouse limb. Group 2 consists of 24 genes expressed in the posterior and the posterior-distal limb. Group 3 consists of 9 genes expressed in the posterior-proximal limb. Group 4 consists of 12 genes expressed in the central portion of the limb. Group 5 consists of 13 genes expressed with multiple spatially distinct domains where one or more expression domain is within the overall Sonic Hedgehog-responsive region in the mouse limb. Group 6 contains a list of all the 205 Sonic Hedgehog target genes (as defined by Vokes et al., 2008). Group 7 contains 147 genes that are not predominately expressed in the Shh responsive region. In other words, this group contains all other gene expression domains: anterior, proximal-anterior, proximal, distal, AER, uniform, weak, and unclear. Group 8 is a variant of Group 1. Group 9 is a variant of Group 4. Group 10 is a variant of Group 7.

We used two methods to generate the binding regions for each group. In method 1, we extracted all of the binding regions specific for genes in each group from list of 396 peak-gene pairs that contain Gli sites for de novo motif discovery. The number of binding regions generated for each group was displayed in Table 2.1. In method 2, we extracted one peak region with the highest MA statistics associated with each gene in each group from list of 656 peak-gene pairs for de novo motif discovery. All analyses were done using the mouse genome assembly mm8.

2.2.2 De Novo Motif Discovery

With the binding regions generated for each group by each method, we performed de novo motif discovery incorporated into CisGenome by running a Gibbs motif sampler three times independently [10]. Gibbs motif sampler is a de novo motif discovery algorithm that searches for enriched sequence patterns in a collection of DNA sequences, which can handle multiple motifs simultaneously [10]. The motifs are assumed to be unknown before the search [10]. Each time, 10 motifs were sampled simultaneously. An initial motif length (L =9, 12, 15) was specified for all motifs at the beginning of the sampling, and the motif lengths were then updated in the sampling procedures [10]. As a consequence, we identified 30 motifs for each group of genes.

A position-specific weight matrix (PWM) was reported for each motif and a motif score was computed as follows:

$$S = \frac{log(n)\sum_{j=A}^{T} p_{ij} log(p_{ij}/q_j)}{W}$$

Here, n is the number of aligned sites that are used to construct the matrix, i.e. $n = n_i = n_{iA}+n_{iC}+n_{iG}+n_{iT}$, where n_{ij} (j = A, C, G, T) is the number of occurrences of nucleotide j at the i-th position of the motif [10]. A pseudocount 0.5 is added to each n_{ij} to avoid zero. $p_{ij} = n_{ij}/n_i$. q_j is the occurrence frequency of nucleotide j in the background sequences [10]. W is the length of the motif.

As a consequence, we discovered 300 motifs totally for binding regions generated by method 1. For method 2, we identified 300 motifs totally by de novo motif discovery as well. Motifs with a score less than 1 were considered to have low quality and were excluded from further analysis [10].

2.2.3 Mapping transcription factor binding motif to sequences

When mapping a motif PWM to DNA, background sequences were modeled as a third-order Markov chain [10]. At each position, the likelihood ratio (LR) between the motif model (PWM) and the background model was computed [10]. A site with LR greater than 500 was used to define TFBS. This cutoff represents a balance between sensitivity and specificity of the analysis [10].

For binding regions generated by method 1, we first got 'matched genomic controls' for each group. 'Matched genomic controls' were control regions carefully chosen to match the physical distributions of ChIP-binding regions as described in Ji et al., 2006 [10]. Then, we mapped the motif PWMs discovered by de novo motif discovery to binding regions of each group. As a complementary motif analysis, we also mapped the 525 human and mouse motif matrices from TRANSFAC database to binding regions of each group.

For binding regions generated by method 2, we first got 'matched genomic controls' for each group and mapped the motif PWMs discovered by de novo motif discovery to binding regions of each group. The only difference step was that we did not map 525 human and mouse motif matrices from TRANSFAC database to binding regions of each group.

2.2.4 Examination of motif's relative enrichment levels

In order to identify the key motif that may mediate sequence-specific protein binding, we compared the relative enrichment levels of different motifs in high-quality binding regions versus control genomic regions [10]. Statistics r_1 was defined to characterize relative enrichment levels of a motif in ChIP-binding regions compared to control regions. Assume that n_{1B} counts how many times a motif occurs in ChIP-binding regions, n_{2B} is the total length of non-repeat sequences in ChIP-binding regions, n_{1C} counts how many times the motif occur in control regions and n_{2C} is the total length of non-repeat sequences in control regions [10].

$$r_1 = \frac{n_{1B}/n_{2B}}{n_{1C}/n_{2C}}$$

defines the relative enrichment level of the motif.

For each group, we compared the occurrence rate of motifs in ChIP-binding regions compared to 'matched genomic controls', that is statistics r_1. We chose a motif selection procedure to select enriched motifs by simultaneously requiring $r_1 \geq 2$, motif score $S \geq 1$, number of motif sites $(n_{1B}) \geq max(\frac{1}{5}*(\text{number of genes}), 5)$ (Table 2.1).

For method 1, we applied the motif selection procedure to motifs discovered by de novo motif discovery. This resulted in 11 enriched motifs in group 1, 8 enriched motifs in group 2, 3 enriched motifs in group 4, 4 enriched motifs in group 5, 5

Group no.	Number of Genes	Number of Binding regions	$n_{1B} \geq$ cutoff
Group 1	58	141	12
Group 2	24	68	5
Group 3	9	18	5
Group 4	12	17	5
Group 5	13	38	5
Group 6	205	396	41
Group 7	147	255	30
Group 8	60	149	12
Group 9	14	25	5
Group 10	145	247	29

Table 2.1: The table of gene numbers, number of all binding regions associated with genes in each group and cutoff for number of motif sites in each group for motif selection procedure.

enriched motifs in group 6, 4 enriched motifs in group 7, 10 enriched motifs in group 8, 9 enriched motifs in group 9 and 5 enriched motifs in group 10 discovered by de novo motif discovery. Then, we used TOMTOM motif comparison Tool to visualize their sequence logos with their PWMs as input. Some of motifs reported by the three independent runs had almost the same sequence pattern. These motifs corresponding to the same transcription factor were considered to be redundant and were removed from further analysis. Only one copy with the highest motif score of these redundant motifs was kept after visual inspection of motif logos [10]. After taking unions of motifs enriched in each group, we totally got 23 unique motifs discovered by de novo discovery method.

We applied the same motif selection procedure to TRANSFAC motifs. As a consequence, we found 60 of the 525 motifs enriched in group 1, 96 motifs in enriched group 2, 24 motifs enriched in group 4, 53 motifs enriched in group 5, 19 motifs enriched in group 6, 10 motifs enriched in group 7, 8 motifs enriched in group 8, 72 motifs enriched in group 9 and 74 motifs enriched in group 10. Motifs enriched in different groups can be the same. After taking unions of motifs enriched in each group, we totally had 158 unique motifs from TRANSFAC database enriched in the ten groups.

For method 2, we also applied the same motif selection procedure to motifs discovered by de novo motif discovery and visualized motif logos using TOMTOM motif comparison Tool. After removing redundant motifs, we finally found 3 unique motifs enriched.

2.2.5 Remapping enriched motifs to sequences

We picked out 23 enriched motifs recovered by de novo motif discovery by method 1 and 158 enriched motifs from TRANSFAC database and combined them together as our final enriched motifs. We mapped the PWMs of enriched motifs to binding regions of each group generated by method 1 and set likelihood ratio cutoff to 500. As group 6 contains all the binding regions for the 205 Shh target genes, it is most appropriate for comparing motif's relative enrichment levels between the 10 groups. Thus, we chose the 'matched genomic controls' for group 6 as the common negative control regions for each group. As a result, we had the same denominator of statistics r_1 for each group, n_{1C}/n_{2C}. As some of the motifs occurred very few times in ChIP-binding regions, we added some pseudo-counts to number of motif sites to reduce bias. We define $r_1^\star$ to compare the relative enrichment level of the motif across groups.

$$r_1^\star = \frac{\dfrac{n_{1B} + n_{2B} \times \alpha \times \beta}{(1+\alpha) \times n_{2B}}}{\dfrac{n_{1C} + n_{2C} \times \alpha \times \beta}{(1+\alpha) \times n_{2C}}}$$

Here, we set α to 0.05, $n_{2B} \times \alpha \times \beta \approx 5$. The purpose of construction of new statistics $r_1^\star$ is to set the ratio $r_1^\star$ equal to 1 if $n_{1B} = n_{1C} = 0$. When total length of non-repeat sequences in ChIP-binding regions (n_{2B}) is small, though number of motif sites (n_{1B}) is few, we can still get high enrichment level r_1. Thus, r_1 is not appropriate for comparing motif relative enrichment levels across groups. We compare the relative enrichment of motifs in each group in the log_2 scales. Adding some pseudo-counts can avoid the values of infinity when raw relative enrichment r_1 is 0.

For each group of genes, we chose a new motif selection procedure to filter out motifs with very few motif sites in ChIP-binding regions and select final enriched motifs by simultaneously requiring $r_1^\star \geq 2$, number of motif sites (n_{1B}) $\geq max(\frac{1}{5}*$(number of genes),5). After new motif selection procedure, we filtered out more motifs and finally got 9 enriched motifs discovered by de novo motif discovery and 27 enriched motifs from TRANSFAC database.

Taking unions of the enriched motifs identified from method 1 and method 2, we finally got 36 motifs by method 1 and 1 Hox motif discovered by method 2. We drew a heatmap to compare and visualize the relative enrichment of each motif in each group based on $log_2(r_1^\star)$. Thus, we can check if there are any motifs specifically enriched in one expression group.

2.2.6 Checking quality of Gli motifs

From all the enriched motifs recovered by de novo motif discovery, we chose one Gli motif discovered from group 6 with the highest motif score 5.72. We then mapped the Gli motif matrix to binding regions of each group (Table 2.2). For calculation of log likelihood of motif site, the Gli matrix was compared to a third-order background Markov model and a likelihood ratio ≥ 500 was used as a cutoff to define Gli sites [10]. As for calculation of log likelihood of peak region, the likelihood ratio cutoff was set to 100 for detecting Gli sites. Thus, with a lower cutoff, we could find more motif sites in one peak region. With the Gli consensus-binding pattern, we calculated probability of each motif site in each group.

Position	A	C	G	T
1	15.50	152.50	112.50	33.50
2	50.50	13.50	0.50	249.50
3	12.50	1.50	299.50	0.50
4	1.50	1.50	239.50	71.50
5	0.50	0.50	312.50	0.50
6	15.50	0.50	0.50	297.50
7	0.50	1.50	311.50	0.50
8	0.50	0.50	312.50	0.50
9	0.50	95.50	26.50	191.50
10	17.50	295.50	0.50	0.50

Table 2.2: The motif matrix of Gli motif used for checking Gli quality.

2.2.6.1 Log likelihood of motif site

The log likelihood of motif site is computed as follows:

$$S^{\star} = \sum_{i=1}^{W} log_2(p_i)$$

Here, W is the length of the motif. n is the number of aligned sites that are used to construct the matrix, i.e. $n = n_i = n_{iA} + n_{iC} + n_{iG} + n_{iT}$, where n_{ij} (j = A, C, G, T) is the number of occurrences of nucleotide j at the i-th position of the motif [10]. p_i is the probability of nucleotide j at the i-th position of the motif, $p_{ij} = n_{ij}/n_i$. A pseudocount 0.5 is added to each n_{ij} to avoid zero. The score $S^\star$ is used to measure Gli quality based on motif site, assuming independence between different motif sites within the same peak region. With the log likelihood of each motif site in each group, we plot a boxplot based on score $S^\star$ to compare the relative quality of Gli motif.

2.2.6.2 Log likelihood of motif sites for one peak region

The log likelihood of motif sites for one peak region is computed as follows:

$$\hat{S} = \sum_{K} \sum_{i=1}^{W} log_2(p_i)$$

We suppose there are K motifs in one peak region. For motif k, W is the length of the motif. n is the number of aligned sites that are used to construct the matrix, i.e. $n = n_i = n_{iA} + n_{iC} + n_{iG} + n_{iT}$, where n_{ij} (j = A, C, G, T) is the number of occurrences of nucleotide j at the i-th position of the motif [10]. p_i is the probability of nucleotide j at the i-th position of the motif, $p_{ij} = n_{ij}/n_i$. A pseudocount 0.5 is added to each n_{ij} to avoid zero. The score $\hat{S}$ is the summation of log likelihood of all the motifs in the peak region. $\hat{S}$ is used to measure Gli quality by taking into consideration of correlation between different motif sites in the same peak region. With the log likelihood of motif sites for each peak region in each group, we plot a boxplot based on the score $\hat{S}$ to compare the relative quality of Gli motif.

2.2.6.3 Log likelihood ratio

As a complementary analysis, we also used the log likelihood ratio output from CisGenome to compare the Gli motif quality. The occurrence of Gli sites in conserved genomic segments was modeled as Poisson processes [8]. The null hypothesis H_0 is that the segment is not a Gli-binding region, while the alternative hypothesis H_1 is the segment is a Gli-binding region. If a segment is a Gli-binding region, the occurrence rate of Gli sites was assumed to be λ_1 (site/bp). Under H_1, each Gli site was assumed to be generated from the Gli-binding matrix. In contrast, if a segment is not a Gli-binding region, the occurrence rate of Gli sites was assumed to be λ_0 (site/bp). Under H_0, each site was assumed to be generated from the background Markov model. For each segment, the log likelihood ratio between H_1 and H_0 was computed as:

$$\bar{S} = n \times log_{10}(\frac{\lambda_1}{\lambda_0}) - (\lambda_1 - \lambda_0) \times l \times log_{10}(e) + log_{10}(L_1) - log_{10}(L_0)$$

Here, l is the length of a conserved segment, n is the number of Gli sites in the segment obtained by matrix mapping, L_1 is the probability to generate the sites from the Gli-binding matrix, and L_0 is the probability to generate the sites from background Markov model. The log likelihood ratio $\bar{S}$ is used to measure the Gli enrichment and rank conserved segments [8].

With the log likelihood ratio of each motif site, we plot a boxplot to compare the relative quality of Gli motif.

2.2.6.4 Welch's t-test

Welch's t-test defines the statistic t by the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

where $\bar{X}_i$, $S_i^2$and N_i are the i th sample mean, sample variance and sample size, respectively.

With the log likelihood of each motif site in each group $S^{\star}$, we conducted Welch's t-test on every two groups to compare if there is any difference of quality of Gli motifs between the 10 groups. With the log likelihood of motif sites for each peak region in each group $\hat{S}$, we performed Welch's t-test on every two groups to compare quality of Gli motifs between the 10 groups.

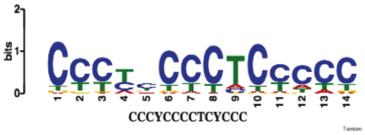
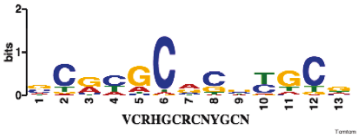
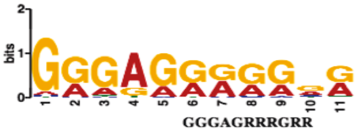
2.3 Results

2.3.1 Motifs identified by de novo motif discovery in each group

After motifs were selected, we listed all the 23 enriched motifs discovered by method 1 (Table 2.3). Gli and Sp1 motifs were found to be highly enriched in each group. We only kept one copy of the Sp1 motifs (Motif 13) with a score of 4.77 and one copy of the Gli motifs (Motif 14) with a score of 5.72. There were also some

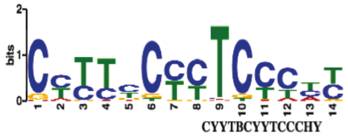
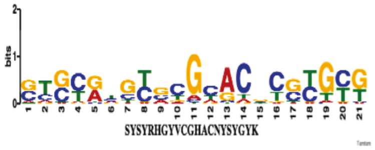
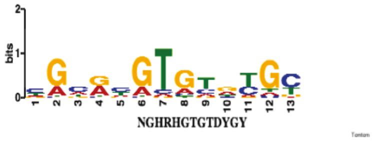
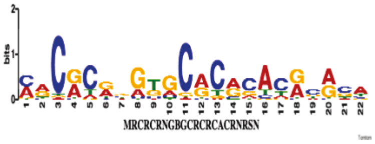
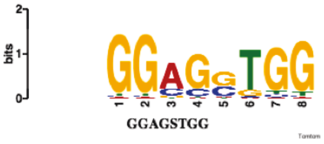
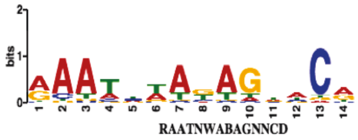
Sp1-like motifs, with GC-rich patterns discovered by de novo motif discovery, such as Motif 1, 3, 4, 10, 11, 21, 22. There were some unknown motifs discovered by de novo motif discovery. We could not find any known motif patterns similar to these motifs.

Table 2.3: Summary of enriched motifs discovered by de novo motif discovery, based all the peak regions associated with genes for each group, number of motif sites (n_{1B}), r_1, motif score S, known motif and motif sequence logos. Motif matrices are represented as sequence logos. Sp1? indicates the motif might be Sp1-like motif but unclear.

Motif	n_{1B}	r_1	Motif score	Known motif	Motif sequence logos
1	74	2.45	2.94	Sp1?	
2	75	2.58	1.53		
3	54	2.03	2.97	Sp1-like	

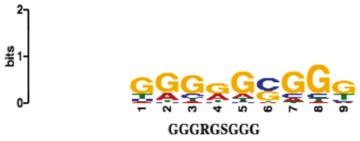
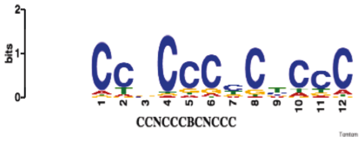
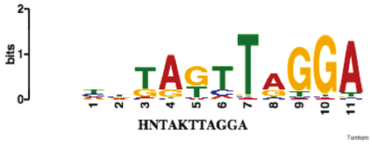
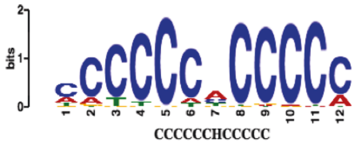
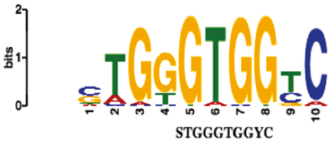
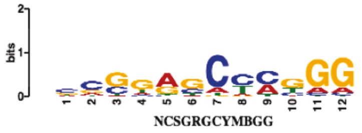
Continued on next page

Table 2.3 – *Continued from previous page*

Motif	n_{1B}	r_1	Motif score	Known motif	Motif sequence logos
4	36	2.05	3.15	Sp1?	
5	36	2.36	1.54		
6	38	3.89	1.48		
7	21	3.30	1.43		
8	7	4.36	2.87		
9	3	3.51	1.23		

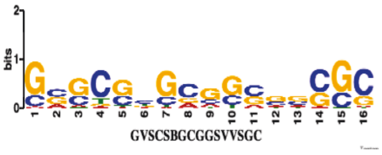
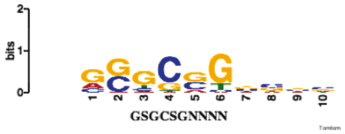
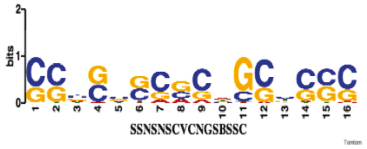
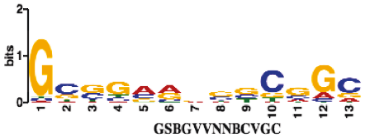
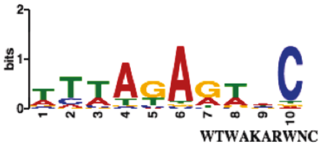
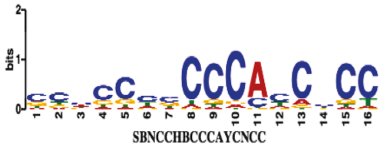
Continued on next page

Table 2.3 – *Continued from previous page*

Motif	n_{1B}	r_1	Motif score	Known motif	Motif sequence logos
10	14	2.38	1.69	Sp1?	
11	21	2.64	2.05	Sp1?	
12	6	2.59	2.15		
13	153	2.27	4.77	Sp1	
14	128	7.89	5.72	Gli	
15	56	2.46	1.66		

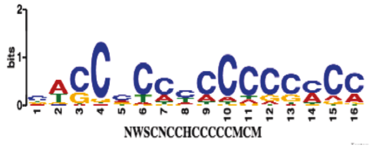
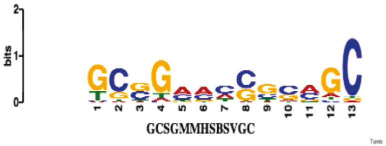
Continued on next page

Table 2.3 – *Continued from previous page*

Motif	n_{1B}	r_1	Motif score	Known motif	Motif sequence logos
16	134	2.18	2.06		
17	83	2.20	1.58		
18	135	2.19	2.19		
19	31	9.26	1.21		
20	3	2.20	1.54		
21	31	4.97	1.62	Sp1?	

Continued on next page

Table 2.3 – *Continued from previous page*

Motif	n_{1B}	r_1	Motif score	Known motif	Motif sequence logos
22	25	5.12	1.82	Sp1?	
23	24	10.64	1.52		

By method 2, after motif selection procedure, we only identified three enriched motifs, Gli, Sp1 and Hox motifs. Gli and Sp1 motifs were still enriched in each group. Hox motif was discovered by de novo motif discovery from peak regions of group 6 and group 10. The one identified from group 10 was kept for further analysis, as the motif score was higher. This Hox motif drew our attention a great deal on account of its function. We will discuss the function of Hox motif later.

For binding regions generated by both methods, Sp1 and Gli motifs are consistently enriched in each group, which means that motifs can be unambiguously identified by de novo motif discovery. For method 2, we lost some peak regions because we just chose one peak associated with each gene. But the analysis might be more informative as the peak regions had the highest moving average statistics. It turned

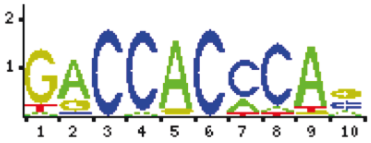
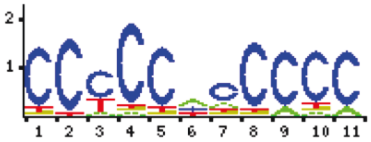
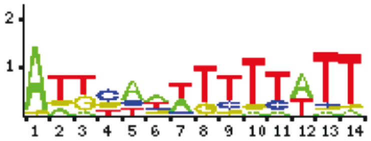
Motif	n_{1B}	r_1	Motif score	Motif sequence logos
Gli	75	7.23	3.81	
Sp1	134	2.33	2.97	
Hox	36	3.05	2.13	

Table 2.4: Summary of enriched motifs discovered by de novo motif discovery, based one peak region with highest MA statistics associated with each gene for each group, known motif, number of motif sites (n_{1B}), r_1, motif score S and motif sequence logos. Motif matrices are represented as sequence logos.

out that Hox motif was only discovered by method 2. This might result from the fact that the 396 peak-gene pairs in method 1 were binding regions containing Gli motifs, while Hox motif might be enriched in binding regions without Gli motifs. Several unknown motifs were discovered by method 1. Though unknown, some of these motifs might be functional elements.

2.3.2 Analysis of a novel motif in Gli binding regions

By insight of the sequence logo of the motif (Table 2.3), Motif 5 seems uninformative. Although the motif contains not too much information, a few papers have reported motifs similar to Motif 5 which seem functional [11]. The mutation of these motifs affect the ability of the enhancer to turn on the gene. Our hypothesis is that Motif 5 might be a functional element based on the following evidence. In several peak regions, the motif sites of this motif occurred twice. The motif might carry out its function in combinatorial patterns. We aligned the sequences of motif 5 in each peak region to the mouse genome mm10 in UCSC Genome Browser. Browsing in UCSC, we uncovered several important genes near motif sites of this motif, such as Ptch1, Wnt11, Osr2 and Cdk6. The motif sequences are very conserved across species, and some are conserved even in zebrafish genomes. Further investigation will be needed to check whether Motif 5 is a promoter or not.

For example, in peak region 12 (chr5: 3347412-3350438) of group 1 generated by method 1, 4 motif sites were discovered in the forward strand. We aligned the sequences of motif sites to mouse genome mm10 in UCSC Genome Browser (Figure 2.1). We noted that the sequences were of high conservation across species, even conserved in chicken (Figure 2.2). And the sequences were not located in the repeating elements. Cdk6 gene was found to be in the downstream of the sequences.

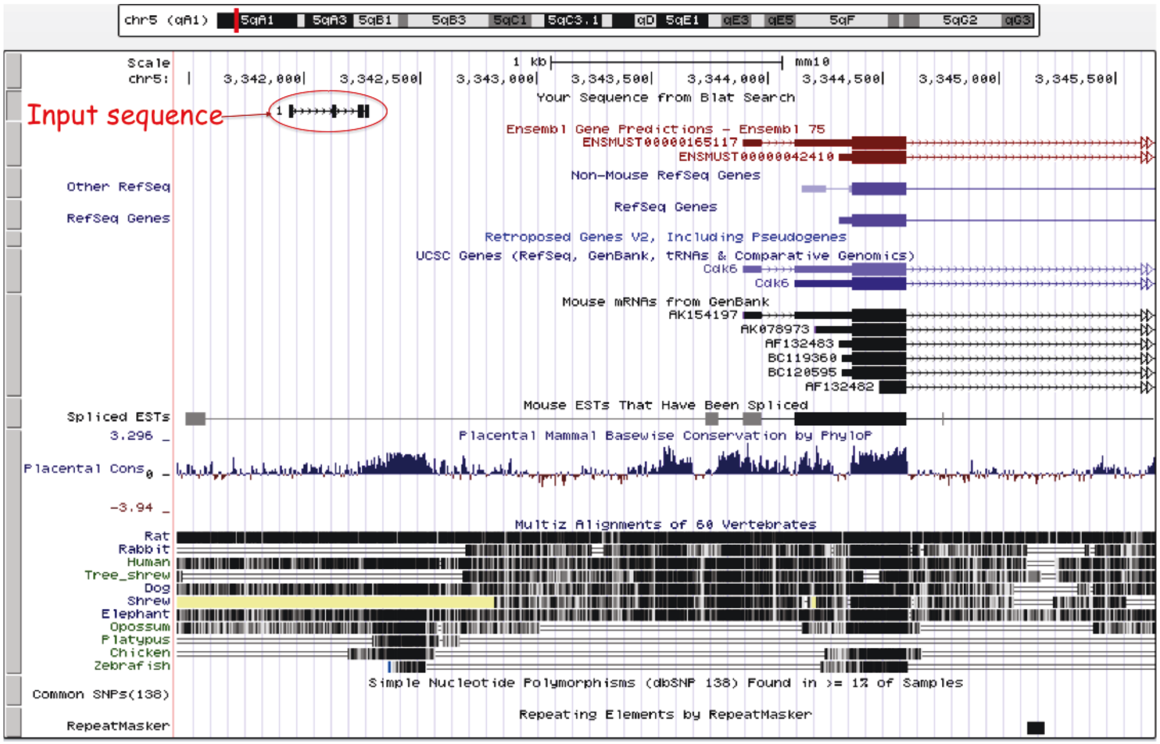


Figure 2.1: Blast results for motif sites of peak region 12 (chr5: 3347412-3350438) in UCSC Genome brower

Gene Cdk6 regulates the synthesis of cyclin-dependent kinase 6 (CDK6). CDK6 is a serine/threonine-protein kinase involved in the control of the cell cycle and differentiation.

2.3.3 Comparison of relative enrichment level of motifs

According to the heatmap that compares the relative enrichment of enriched motifs, the motifs are enriched in the relative group if $log_2(r_1^\star) \geq 1$. As a result, Gli

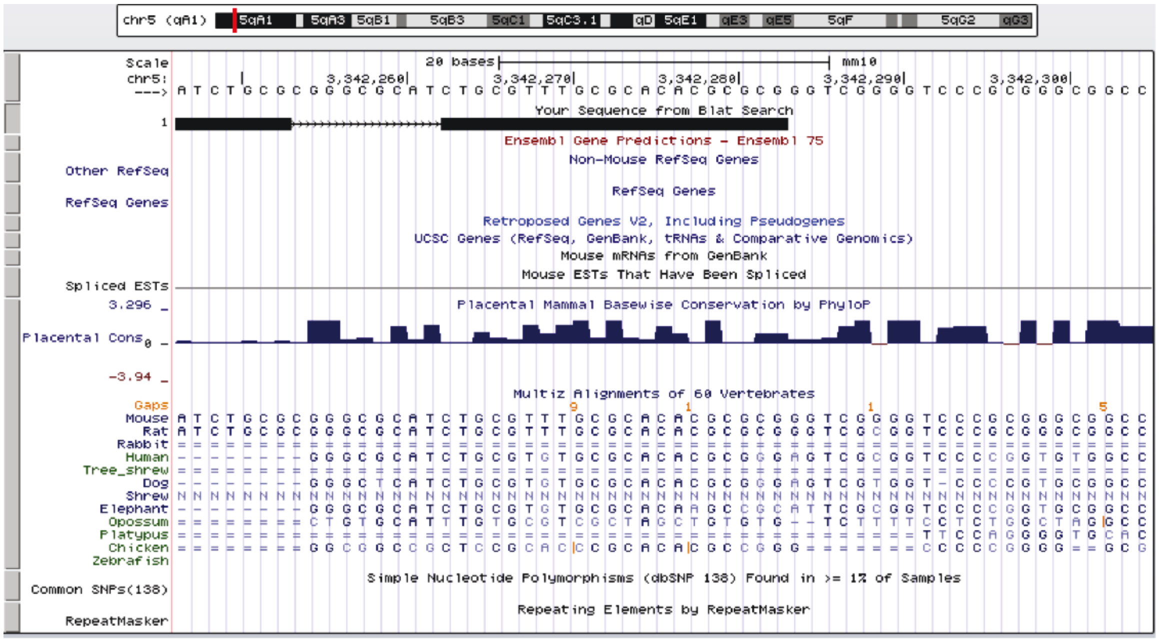

chr5 (qA1)
5qA1
5qA3
5qB1
5qB3
5qC1
5qC3.1
qD
5qE1
qE3
qE5
5qF
5qG2
qG3
Scale
20 bases
mm10
chr5:
3,342,260|
3,342,270|
3,342,280|
3,342,290|
3,342,300|
---> ATCTGCGCGGGCGCATCTGCGTTTGCGCACACGCGCGGGTCGGGGTCCCGCGGGCGGCC
Your Sequence from Blat Search
1
Ensembl Gene Predictions - Ensembl 75
Non-Mouse RefSeq Genes
Other RefSeq
RefSeq Genes
RefSeq Genes
Retroposed Genes V2, Including Pseudogenes
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced
Spliced ESTs
3.296
Placental Mammal Basewise Conservation by PhyloP
Placental Cons
0
-3.94
Multiz Alignments of 60 Vertebrates
Gaps
9
1
1
5
Mouse ATCTGCGCGGGCGCATCTGCGTTTGCGCACACGCGCGGGTCGGGGTCCCGCGGGCGGCC
Rat ATCTGCGCGGGCGCATCTGCGTTTGCGCACACGCGCGGGTCGCGGTCCCGCGGGCGGCC
Rabbit ===
Human --------GGGCGCATCTGCGTGTGCGCACACGCGGGAGTCGCGGTCCCCGGTGTGGCC
Tree_shrew ===
Dog --------GGGCTCATCTGCGTGTGCGCACACGCGGGAGTCGTGGT-CCCCGTGCGGCC
Shrew NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Elephant --------GGGCGCATCTGCGTGTGCGCACAAGCCGCATTCGCGGTCCCCGGTGCGGCC
Opossum ========CTGTGCATTTGTGCGTCGCTAGCTGTGTG--TCTTTTCCTCTGGCTAGGCC
Platypus ==TTCCAGGGGTGCAC
Chicken ========GGCGGCCGCTCCGCACCCGCACACGCCGGG=======CCCCCGGGG==GCG
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
Common SNPs(138)
Repeating Elements by RepeatMasker
RepeatMasker

Figure 2.2: Blast results across species for the fourth motif site of peak region 12 (chr5: 3347412-3350438) in UCSC Genome brower

motif is enriched in each group, except for group 3 and group 4. This might be result from the few peak regions in group 3 and group 4, thus some of the motifs might not pass the signal-noise-ratio threshold. ZIC1, ZIC2 and ZIC3 motifs, are all Gli motifs, enriched in group 1, group 2, group 6, group 7, group 8, and group 10. SP1 motif is enriched in group 1, group 3, group 4, group 6, group 7 and group 10. STAT1 motif is enriched group 1, group 6 and group 10. AP2, Motif 23, Motif 19, Motif 22 (SP1-like) and Motif 21 (SP1-like) motifs are enriched in group 4 and group 9. RREB1 motif is enriched in group 2 and group 8. E2F1 motif is enriched in group 2, group 4 and group 8. E2F motif is enriched in group 1, group 4, group 6, group 7, group 9 and group 10. Motif 1, which is an SP1-like motif, is enriched in group 1 and group 4.

There are also some motifs that are specifically enriched in one group. AP2GAMMA, AP2ALPHA, ETF, YY1 motifs are only enriched in group 9. PAX4 is only enriched in group 3. ELK1 and ZF5, PAX2, NRF1, Motif 7 and Motif 6 are only enriched in group 2.

As to the Hox motif, it is not enriched in any of the ten groups. It is consistent with our results as we did not discover the Hox motif in all of the peak regions by method 1. The enrichment levels of Hox motif are relatively high in group 7 and group 10, compared with other groups.

Thus, we have identified some motifs specifically enriched in some of the groups. These motifs might be specifically associated with the relative targeted genes.

2.3.4 There is no statistically significant difference of Gli quality among different groups.

From the boxplot (Figure 2.4), we can easily see that the mean of log likelihood score $S^{\star}$ of group 4 is very low compared to the other groups. The boxplot of log likelihood ratio score $\bar{S}$ (Figure 2.4) is consistent with the boxplot of log likelihood score $S^{\star}$. We reported the p-value of Welchs t-test on every two groups based on log likelihood of each motif site in each group (Table 2.5). There is significant difference of Gli quality between group 4 and group 5 (p-value 0.04), group 4 and group 6 (p-value 0.044), group 4 and group 7 (p-value 0.034). However, according to the histogram

(Figure 2.5), we can see that the t-tests of Gli quality are dependent thus we can not make inference based on raw p-value. When the tests are correlated with each other, p-values are not uniformly distributed. Having a total of 100 hypotheses, the Type I error with Bonferroni correction is 0.0005 (i.e. 0.05/100). Thus, by assuring family wise error rate (FWER) less than 0.05, the probability of making even one type I error in the family is controlled at level 0.05. But after Bonferroni correction, there is no significant difference of Gli quality between different groups.

	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9	Group10
Group1	1.00	0.65	0.78	0.08	0.37	0.22	0.09	0.65	0.76	0.13
Group2	0.65	1.00	0.94	0.06	0.66	0.67	0.42	0.93	0.57	0.50
Group3	0.78	0.94	1.00	0.14	0.87	0.93	0.81	0.91	0.66	0.84
Group4	0.08	0.06	0.14	1.00	0.04	0.04	0.03	0.06	0.17	0.04
Group5	0.37	0.66	0.87	0.04	1.00	0.85	0.87	0.56	0.39	0.96
Group6	0.22	0.67	0.93	0.04	0.85	1.00	0.43	0.47	0.38	0.60
Group7	0.09	0.42	0.81	0.03	0.87	0.43	1.00	0.22	0.27	0.82
Group8	0.65	0.93	0.91	0.06	0.56	0.47	0.22	1.00	0.58	0.29
Group9	0.76	0.57	0.66	0.17	0.39	0.38	0.27	0.58	1.00	0.30
Group10	0.13	0.50	0.84	0.04	0.96	0.60	0.82	0.29	0.30	1.00

Table 2.5: P-value output from Welch's t-test on every two groups.

Considering of correlation between motif sites, we reported the p-value of Welch's t-test on every two groups based on log likelihood of motif sites in each peak region in each group $\hat{S}$ (Table 2.6). From the table 2.6, we can see that there is no statistically significant difference of Gli quality between the ten groups. The boxplot (Figure 2.7) also shows that there is not much difference between the mean of log likelihood of motif sites in each peak region in each group $\hat{S}$. According to the histogram (Figure 2.6), p-values are not uniformly distributed thus Bonferroni correction is needed. After Bonferroni correction, there is still no statistically significant difference of Gli quality between the ten groups.

Thus, we conclude that there is no statistically significant difference of quality between Gli motifs in Gli binding regions associated with genes expressed in the Shh-responsive region and Gli motifs in Gli binding regions associated with genes that are not expressed in the Shh-responsive region. In addition, there is no statistically significant difference of quality of Gli motifs in Gli binding regions associated with genes expressed in individual expression groups.

	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9	Group10
Group1	1.00	0.56	0.15	0.47	0.81	0.36	0.21	0.98	0.50	0.18
Group2	0.56	1.00	0.40	0.37	0.48	0.99	0.78	0.54	0.34	0.73
Group3	0.15	0.40	1.00	0.24	0.14	0.30	0.42	0.14	0.15	0.46
Group4	0.47	0.37	0.24	1.00	0.52	0.36	0.33	0.47	0.76	0.32
Group5	0.81	0.48	0.14	0.52	1.00	0.36	0.25	0.82	0.61	0.23
Group6	0.36	0.99	0.30	0.36	0.36	1.00	0.56	0.32	0.30	0.47
Group7	0.21	0.78	0.42	0.33	0.25	0.56	1.00	0.18	0.25	0.89
Group8	0.98	0.54	0.14	0.47	0.82	0.32	0.18	1.00	0.50	0.15
Group9	0.50	0.34	0.15	0.76	0.61	0.30	0.25	0.50	1.00	0.23
Group10	0.18	0.73	0.46	0.32	0.23	0.47	0.89	0.15	0.23	1.00

Table 2.6: P-value output from Welch's t-test on every two groups adjusting for correlation.

2.4 Summary and Discussion

2.4.1 Summary

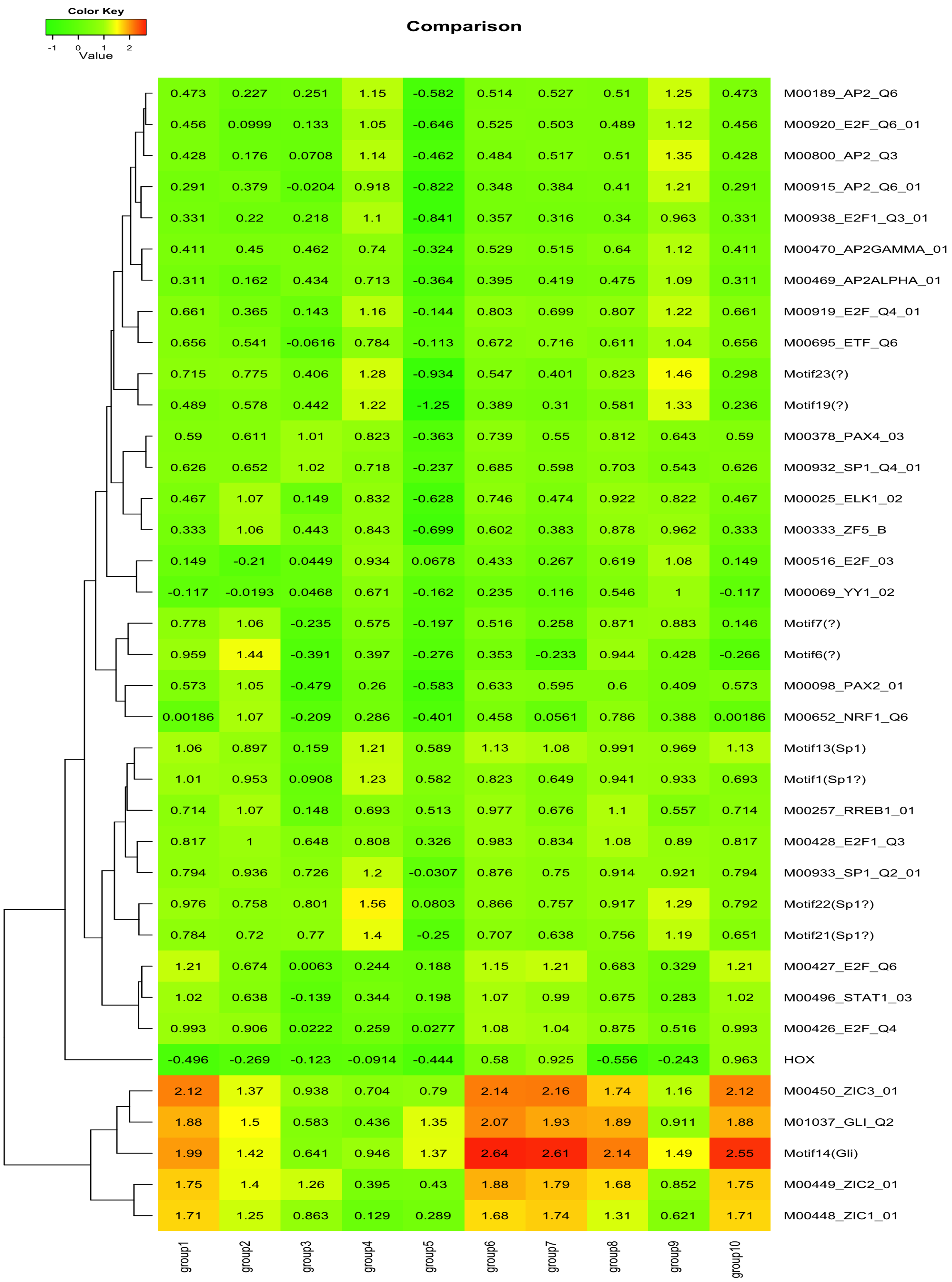
As a result, we found that Gli and Sp1 motifs were highly enriched in all Shh responsive regions (Figure 2.8). There were some motifs enriched specifically in individual categories. There were also some Sp1-like motifs, with GC-rich patterns enriched in all Shh responsive regions, such as Motif 1, 3, 4. GC-rich pattern motifs were also found in multiple domains, such as Motif 10, 11. We discovered some unknown motifs (Figure 2.8) in different gene expression categories, such as Motif 5. Motif 5, enriched in all Shh responsive regions, might be a functional element. There is no statistically significant difference of Gli quality between Gli motifs in Gli binding regions associated with genes expressed in the Shh-responsive region and Gli motifs in Gli binding regions associated with genes that are not expressed in the Shh-responsive region. Furthermore, there is no statistically significant difference of quality of Gli motifs in Gli binding regions associated with genes expressed in individual expression groups. The quality of Gli motif might not have influence on the expression of genes in different domains.

2.4.2 Discussion

In our study, Hox motif is found in genes that are not predominately expressed in the Shh-responsive region in the limb. This could suggest that Hox genes may not function within GBRs to regulate Shh target genes. Genetic studies have indicated that 5'HoxD complex have been implicated in the regulation of digit identity and Shh [9]. 5'HoxD not only plays an important function in the onset of Shh expression but also becomes targets of Shh regulation [9]. HOXA and/or HOXD proteins have been confirmed to be required for Shh transcription [12]. Under the control of HOX proteins, confined expression of Sonic hedgehog (Shh) at the posterior margin of developing early limb buds determines the anterior to posterior (AP) polarity of the limb [12]. We also found that enrichment level of HoxA motif was relatively high in the Shh-responsive region in the limb. To study how Hox may function in Gli binding regions, we need to do further experiments to figure it out. We could mutate Hox gene and find if this mutation can influence the development of limb buds. Our collaborator is planning to do some further experiments on Hox motif. They have cloned about 40 of the GBRs into reporter vectors to test for activity, silencer or enhancer. And, they will mutate some specific sites of Hox motif to see if mutations change reporter expression.

Gli Binding regions contain GC rich sequences that could possibly bind other factors like SP1 or Kruppel like factors (KLFs), which are both found in the limb bud. In the four categories, we found some unknown motifs enriched in each category, except for genes expressed in the posterior-proximal limb. Motif 8 is enriched in genes expressed in the central portion of the limb, which contains a TGG sequence. It is not possible that this motif is a Gli motif but is cut off by the algorithm. The motif pattern is not similar to Gli motif. By running Glibbs motif sampler three times independently, Gli motif should be identified if it is enriched in the binding regions. As there are too few peak regions in this group, so it is possible that Gli motif could not be identified by de novo motif discovery. Some of the motifs might not pass the signal-noise-ratio threshold. To improve analysis, we could generate more peak regions in each category, or develop a new algorithm that can detect some weak motifs even with few peak regions.

In order to check whether Motif 5 is a function element or not, we will map the PWM of Motif 5 to mouse whole genome. With the motif sites identified, we can count how many motif sites are located within 1kb upstream of transcription start site (TSS), 1 kb downstream of transcription end site (TES), Intragene and Intergene. We will cluster the motif sites and repeat the same analyses on the mouse genome. If we can find a strong correlation between the clustered sites and promoters, Motif 5 is highly likely to be a functional promoter element [2].



	group1	group2	group3	group4	group5	group6	group7	group8	group9	group10
M00189_AP2_Q6	0.473	0.227	0.251	1.15	-0.582	0.514	0.527	0.51	1.25	0.473
M00920_E2F_Q6_01	0.456	0.0999	0.133	1.05	-0.646	0.525	0.503	0.489	1.12	0.456
M00800_AP2_Q3	0.428	0.176	0.0708	1.14	-0.462	0.484	0.517	0.51	1.35	0.428
M00915_AP2_Q6_01	0.291	0.379	-0.0204	0.918	-0.822	0.348	0.384	0.41	1.21	0.291
M00938_E2F1_Q3_01	0.331	0.22	0.218	1.1	-0.841	0.357	0.316	0.34	0.963	0.331
M00470_AP2GAMMA_01	0.411	0.45	0.462	0.74	-0.324	0.529	0.515	0.64	1.12	0.411
M00469_AP2ALPHA_01	0.311	0.162	0.434	0.713	-0.364	0.395	0.419	0.475	1.09	0.311
M00919_E2F_Q4_01	0.661	0.365	0.143	1.16	-0.144	0.803	0.699	0.807	1.22	0.661
M00695_ETF_Q6	0.656	0.541	-0.0616	0.784	-0.113	0.672	0.716	0.611	1.04	0.656
Motif23(?)	0.715	0.775	0.406	1.28	-0.934	0.547	0.401	0.823	1.46	0.298
Motif19(?)	0.489	0.578	0.442	1.22	-1.25	0.389	0.31	0.581	1.33	0.236
M00378_PAX4_03	0.59	0.611	1.01	0.823	-0.363	0.739	0.55	0.812	0.643	0.59
M00932_SP1_Q4_01	0.626	0.652	1.02	0.718	-0.237	0.685	0.598	0.703	0.543	0.626
M00025_ELK1_02	0.467	1.07	0.149	0.832	-0.628	0.746	0.474	0.922	0.822	0.467
M00333_ZF5_B	0.333	1.06	0.443	0.843	-0.699	0.602	0.383	0.878	0.962	0.333
M00516_E2F_03	0.149	-0.21	0.0449	0.934	0.0678	0.433	0.267	0.619	1.08	0.149
M00069_YY1_02	-0.117	-0.0193	0.0468	0.671	-0.162	0.235	0.116	0.546	1	-0.117
Motif7(?)	0.778	1.06	-0.235	0.575	-0.197	0.516	0.258	0.871	0.883	0.146
Motif6(?)	0.959	1.44	-0.391	0.397	-0.276	0.353	-0.233	0.944	0.428	-0.266
M00098_PAX2_01	0.573	1.05	-0.479	0.26	-0.583	0.633	0.595	0.6	0.409	0.573
M00652_NRF1_Q6	0.00186	1.07	-0.209	0.286	-0.401	0.458	0.0561	0.786	0.388	0.00186
Motif13(Sp1)	1.06	0.897	0.159	1.21	0.589	1.13	1.08	0.991	0.969	1.13
Motif1(Sp1?)	1.01	0.953	0.0908	1.23	0.582	0.823	0.649	0.941	0.933	0.693
M00257_RREB1_01	0.714	1.07	0.148	0.693	0.513	0.977	0.676	1.1	0.557	0.714
M00428_E2F1_Q3	0.817	1	0.648	0.808	0.326	0.983	0.834	1.08	0.89	0.817
M00933_SP1_Q2_01	0.794	0.936	0.726	1.2	-0.0307	0.876	0.75	0.914	0.921	0.794
Motif22(Sp1?)	0.976	0.758	0.801	1.56	0.0803	0.866	0.757	0.917	1.29	0.792
Motif21(Sp1?)	0.784	0.72	0.77	1.4	-0.25	0.707	0.638	0.756	1.19	0.651
M00427_E2F_Q6	1.21	0.674	0.0063	0.244	0.188	1.15	1.21	0.683	0.329	1.21
M00496_STAT1_03	1.02	0.638	-0.139	0.344	0.198	1.07	0.99	0.675	0.283	1.02
M00426_E2F_Q4	0.993	0.906	0.0222	0.259	0.0277	1.08	1.04	0.875	0.516	0.993
HOX	-0.496	-0.269	-0.123	-0.0914	-0.444	0.58	0.925	-0.556	-0.243	0.963
M00450_ZIC3_01	2.12	1.37	0.938	0.704	0.79	2.14	2.16	1.74	1.16	2.12
M01037_GLI_Q2	1.88	1.5	0.583	0.436	1.35	2.07	1.93	1.89	0.911	1.88
Motif14(Gli)	1.99	1.42	0.641	0.946	1.37	2.64	2.61	2.14	1.49	2.55
M00449_ZIC2_01	1.75	1.4	1.26	0.395	0.43	1.88	1.79	1.68	0.852	1.75
M00448_ZIC1_01	1.71	1.25	0.863	0.129	0.289	1.68	1.74	1.31	0.621	1.71

Figure 2.3: Heatmap of motifs' enrichment levels for each group. 37 enriched motifs were compared based on $log_2(r^{\star})$.

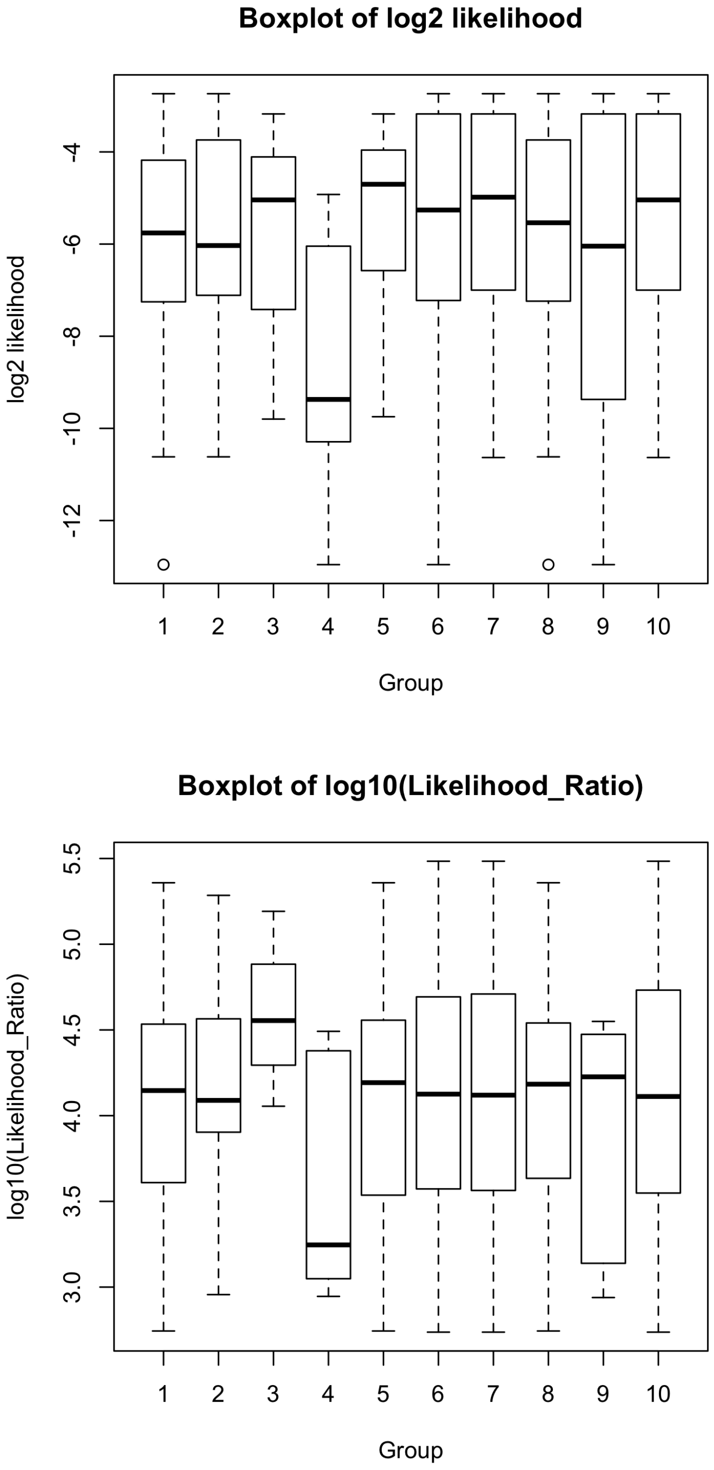


Figure 2.4: The boxplot of log likelihood score $S^{\star}$ and log likelihood ratio $\bar{S}$ across groups.

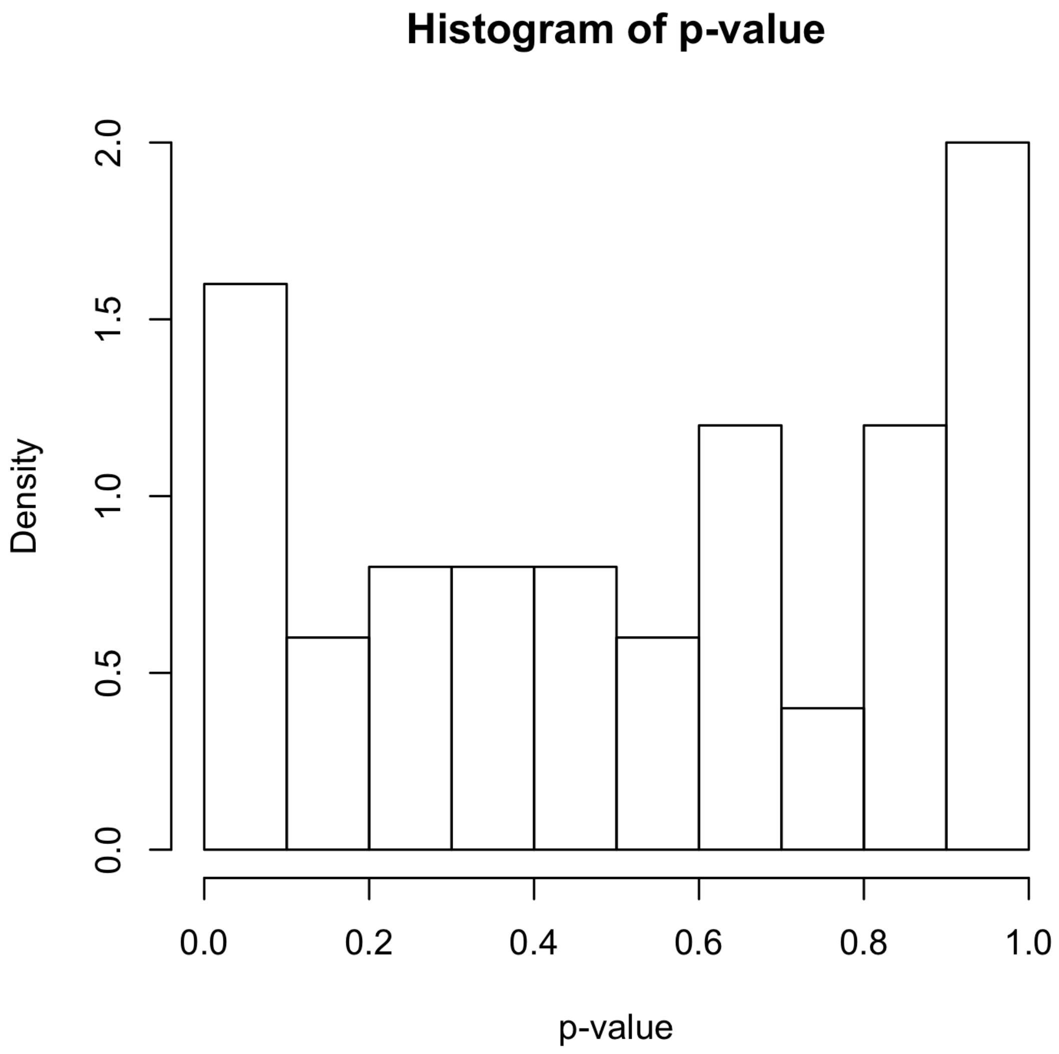


Figure 2.5: Histogram of raw p-value output from Welch's t-test on every two groups.

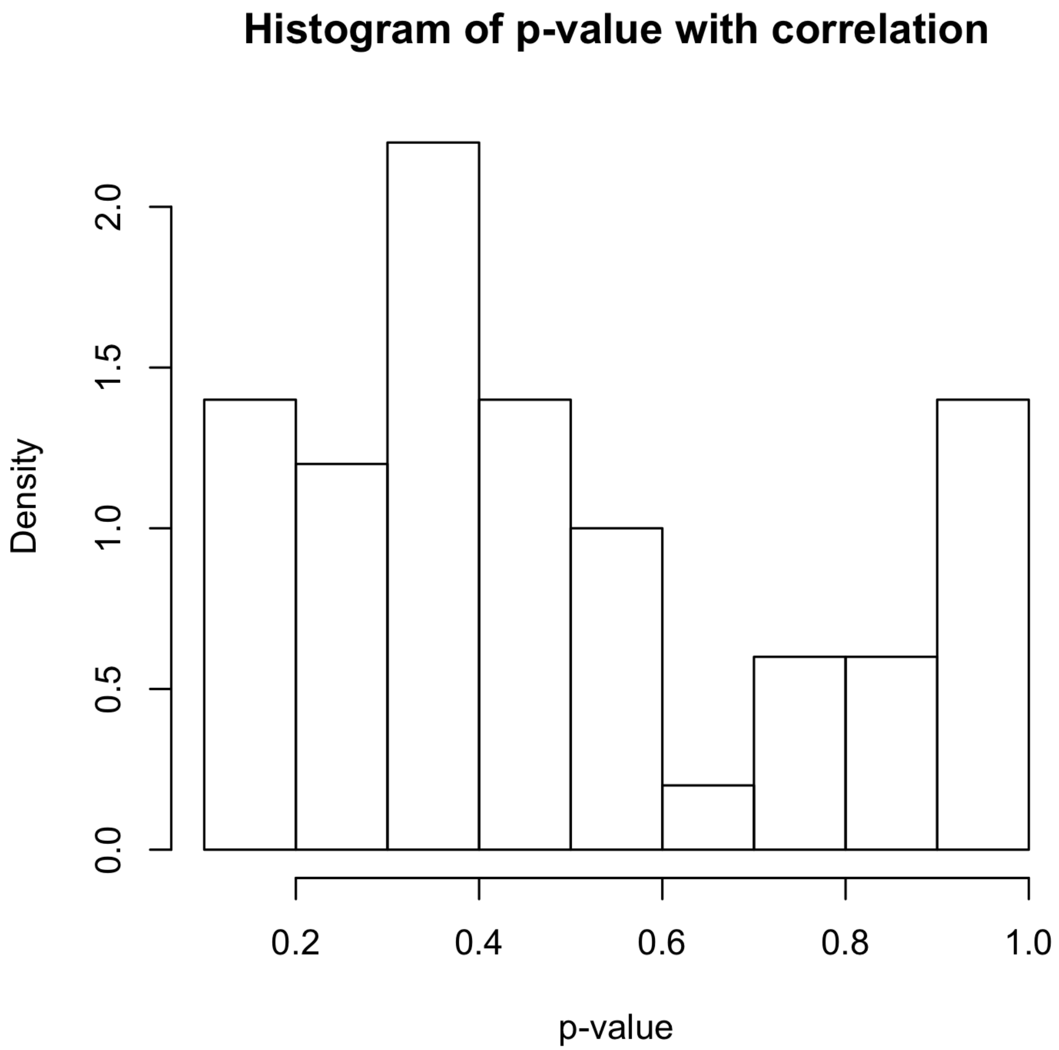


Figure 2.6: Histogram of raw p-value output from Welch's t-test on every two groups adjusting for correlation.

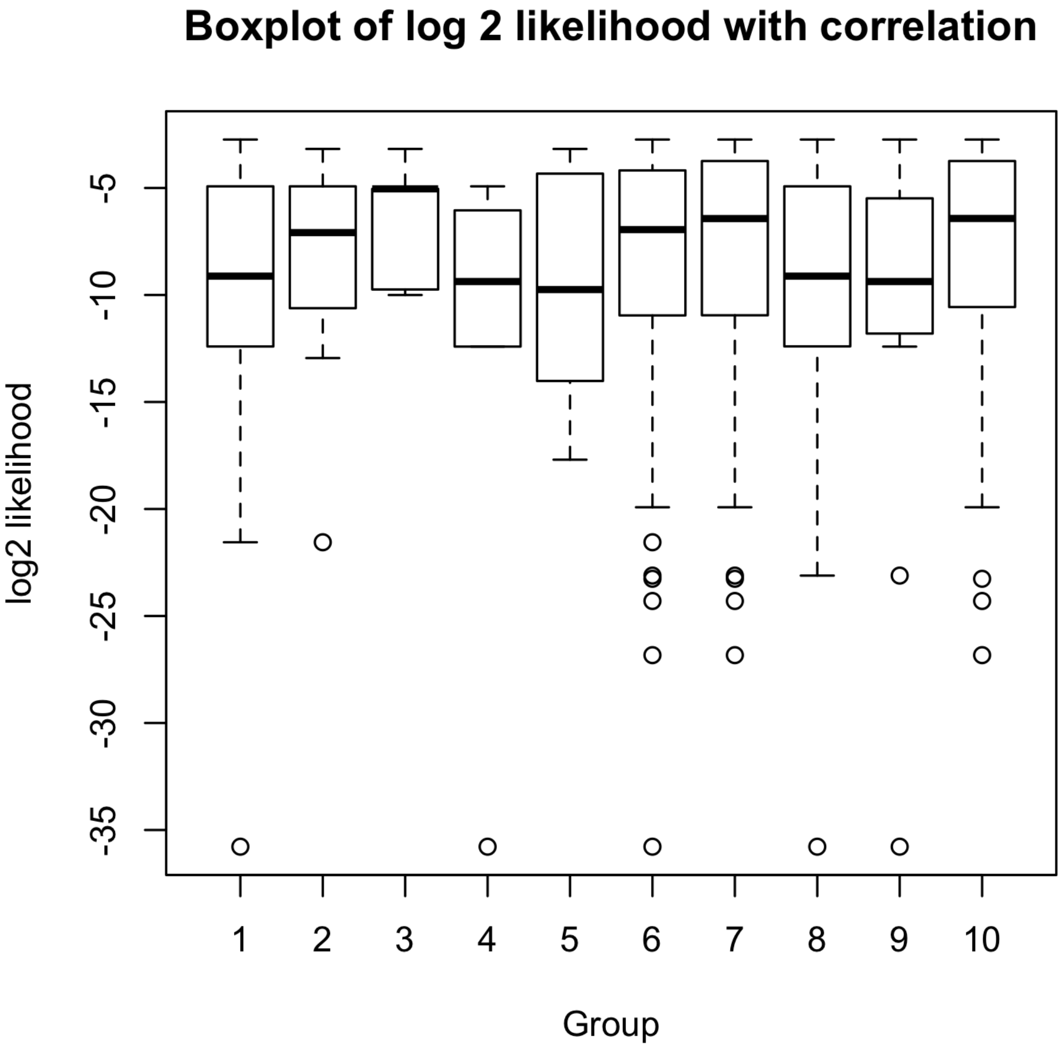


Figure 2.7: The boxplot of log likelihood score $S^{\star}$ across groups adjusting for correlation.

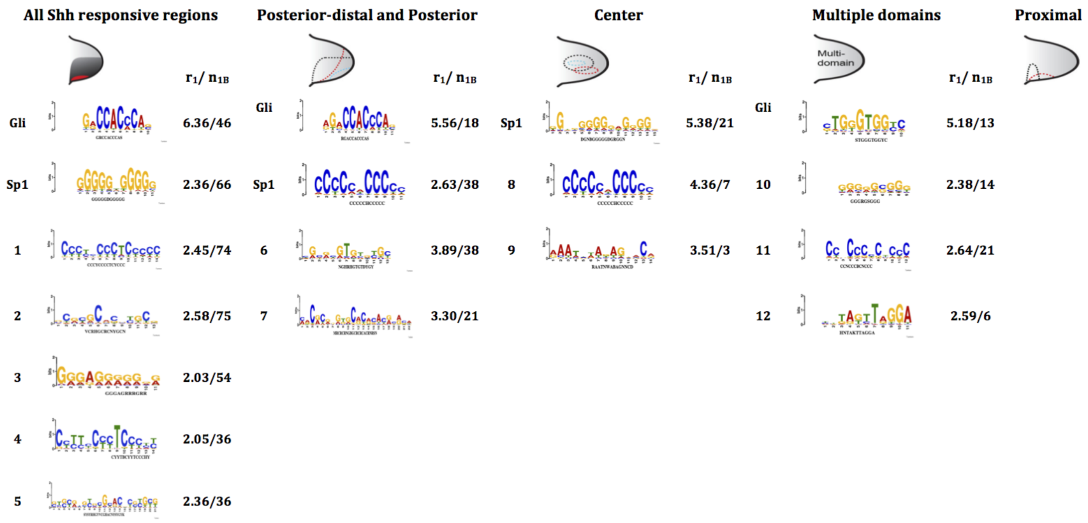


Figure 2.8: Summary of motifs enriched in each category, with motif's enrichment level r_1, number of motif sites in ChIP-binding regions n_{1B}. The motif number is the same as in table 2.3. The cartoons of E10.5 forelimbs showing the observed expression patterns of each category are from Dr.Vokes's laboratory.

Chapter 3

Antibody-specific motif analysis with application to human MYC

3.1 Introduction

Myc protein is a transcription factor encoded by the c-MYC gene (thereafter termed MYC), which regulates around 15% of genes in the human genome [13]. Myc is a helix-loop-helix leucine zipper transcription factor, which forms a heterodimer with MYC-associated factor X (Max). The Myc and Max heterodimer binds to the DNA consensus sequence, Enhancer Box (E-box) having the sequence 5'-CACG/ATG-3' [13, 14]. Upon binding, Myc/Max recruits cofactors that regulate the transcription of distinct genes which are involved in cell cycle progression, differentiation and apoptosis [13, 14]. Myc has the ability to activate and repress gene transcription. Elevated expression of Myc transcription factor is frequently found in 70% of human tumors [13]. Myc/Max dimers associate with other transcription factors to repress gene transcription, such as Miz-1 or NF-Y [15]. Myc plays an important role in stem cell pluripotency, self-renewal and induction of adult cells back to pluripotent state [14]. The role of Myc is found to be cell type and species independent [14].

Chromatin immunoprecipitation (ChIP) was performed using two different anti-Myc antibodies in human P493-6 B cells, a model of Burkitt lymphoma (BL) that had an Epstein-Barr virus genome and a tetracycline (tet)-repressible human MYC transgene [14]. The choice of model of human P493-6 B cells is because human B lymphoma cells could bear inducible MYC [14]. Burkitt lymphoma is an outstanding example for MYC overexpression due to a chromosomal translocation involving the c-MYC gene [13]. After removal of tetracycline, Myc protein is highly induced [14]. As a result, resting P493-6 cells are recruited into the active cell cycle [14]. The two different anti-Myc antibodies are the Santa Cruz (SC) anti-N-terminal Myc antibody and the Epitomics (Epit) monoclonal anti-N-terminal Myc antibody [14]. A sensitive and specific antibody will give a high level of enrichment compared with the background, which makes it easier to detect binding events [4]. We sought to unravel if the binding targets of Myc transcription factor are antibody-specific. The interactions of Myc with different antibodies are distinct as they have different binding sites. The different binding patterns can influence the affinity of Myc with DNA motifs. With different antibodies, the Myc-antibody complexes might have different structures, thus their preference of DNA motifs might be different. If there exist antibody-specific DNA motifs of transcription factors, we can effectively select specific antibody for ChIP experiments to study the mechanism of gene regulations.

Thus, the goal of our study is to identify antibody-specific motifs with application to Myc transcription factor by using high-throughput genomic approaches.

3.2 Methods

3.2.1 Data preparation

We collected ChIP-chip data for Myc TF from Gene Expression Omnibus (GEO, Human, GEO accession no: GSE32220). The ChIP-chip datasets were generated using Affymetrix human promoter 1.0R array platform [10]. The ChIP-chip data were analyzed using CisGenome. All analyses were done using the human genome assembly hg19.

3.2.2 General data analysis protocol

Cisgenome first loaded the raw data and then did the quantile normalization, finally exported the intensities of each data [10]. After quantile normalization, we applied TileMapv2 Moving Avarage (MA) algorithm for ChIP-chip peak detection im-

Data	Species	Type	Platform	Samples	Sources
P493 B cells	Human	ChIP-chip (Santa Cruz antibody)	Affymetrix Human Promoter 1.0R Arrays	6 (3 IP vs.3 IgG control)	GEO (GSE32220)
P493 B cells	Human	ChIP-chip (Epit antibody)	Affymetrix Human Promoter 1.0R Arrays	4 (2 IP vs.2 IgG control)	GEO (GSE32220)

Table 3.1: ChIP-chip Data Used in the Analysis

plemented in CisGenome to define potential protein-binding regions for the ChIP-chip datasets (Myc) [10]. Here we have two antibodies: Epitomics (Epit) monoclonal anti-N-terminal Myc antibody and Santa Cruz(SC) anti-N-terminal Myc antibody [14]. In human P493-6 B cells, the Santa Cruz (SC) anti-N-terminal Myc antibody revealed Myc binding to 2452 regions. Using the Epitomics (Epit) monoclonal anti-N-terminal Myc antibody, 1957 Myc binding regions were identified. We took union of the peak lists from two antibodies together and got 3507 binding regions. Then, we output the average probe intensities in each peak region of the 8 samples from Cisgenome, 3 IPs from SC antibody, 2 IPs from Epit antibody and 3 IgG controls. With the average probe intensities of the samples, we can check if there is any statistically significant difference between the binding intensity of probe with ChIP sequence immunoprecipitated by different antibodies.

3.2.3 Comparison of probe intensity between two antibodies using limma

Limma is a package from Bioconductor for the analysis of gene expression microarray data, especially the use of linear models for analyzing designed experiments and the assessment of differential expression [16]. With the average probe intensities of the IP samples, we can check difference of probe intensity between two antibodies by limma package from Bioconductor.

3.2.3.1 Checking normalization

First, we plot a scatter plot for each replicate of the sample to visualize if the data were normalized. If the data were not well normalized, the probe intensities across samples were not comparable thus differentiation analysis by limma would not make sense.

3.2.3.2 Linear model

We used limma package to check if there is any difference of probe intensity between SC antibody and Epit antibody groups. We fitted a multiple linear models by weighted or generalized least squares and empirical Bayes methods using the input data. Our design for the linear model is:

$$y = \beta_0 + \beta_1 * x$$

Here x is a dummy variable, taking values of 0 or 1. $x = 0$ indicates the Epit antibody group, while $x = 1$ indicates the SC antibody group. y indicates average probe intensities for each IP sample. β_1 characterizes the difference of average probe intensities between the two antibodies. β_0 is the average probe intensity of IP samples from Epit antibody group.

And we checked difference of average probe intensity by performing Wald's test. If the coefficient of different groups, i.e. β_1 is significant different from 0, we consider there might be statistically significant difference of average probe intensity between the two antibodies. Our significance level is set at 0.05.

3.2.3.3 Generating new groups

Here, adjusted p-value (i.e. q-value) is used to check if there is any difference of average probe intensity between the two antibodies. When adjusted p-value is larger than 0.05, we conclude that there is no statistically significant difference between the two antibodies. When adjusted p-value is less than 0.05, we conclude that there is statistically significant difference between the two antibodies. Based on adjusted p-value and t statistics, we divided the 3507 peak regions into five groups:

Myc_Group 1: t-statistic> 0, q-value< 0.05 (significant difference, SC antibody strong binding)

Myc_Group 2: t-statistic< 0, q-value< 0.05 (significant difference, Epit antibody strong binding)

Myc_Group 3: q-value$>$ 0.5 (no difference)

Myc_Group 4: t-statistic$>$ 0, 0.05 $<$q-value$<$ 0.5 (grey area)

Myc_Group 5: t-statistic$<$ 0, 0.05 $<$q-value$<$ 0.5 (grey area)

Myc_Group 1 and Myc_Group 2 represent regions having significant difference of binding intensity between the two antibodies. Regions in Myc_Group 1 have a higher binding intensity with SC antibody, while regions in Myc_Group 2 have a higher binding intensity with Epit antibody. Myc_Group 3 represents regions that there is no significant difference of binding intensity between the two antibodies. Myc_Group 4 and Myc_Group 5 represent regions of grey area, which are unclear.

3.2.3.4 De novo motif discovery

For the five new groups of peak regions, enriched sequence patterns were identified through de novo motif discovery. Each time, 20 motifs were sampled simultaneously. An initial motif length (L =9, 12, 15) was specified for all motifs at the beginning of the sampling, and the motif lengths were then adjusted during the sampling procedures. A position-specific weight matrix (PWM) and motif score were reported for each motif. As a consequence, we identified 60 motifs for each group.

3.2.3.5 Mapping transcription factor binding motif to sequences

In order to identify the key motif that may mediate sequence-specific protein binding, we compared different motifs' relative enrichment levels in ChIP-binding regions versus control genomic regions. With the 300 motifs discovered in the previous step and 525 human and mouse motif matrices from TRANSFAC database, we mapped the PWM of total 825 motifs to each group of binding regions and set likelihood ratio cutoff to 500.

$$r_1 = \frac{n_{1B}/n_{2B}}{n_{1C}/n_{2C}}$$

defines the relative enrichment level of the motif. It is the same with detailed description in Chapter 2.

For each group, we compared the occurrence rate of Myc-binding regions compared to 'matched genomic controls', i.e. statistics r_1. We chose a cutoff to define enriched motifs by simultaneously requiring motif score $S \geq 1$, $r_1 \geq 2, n_{1B} \geq max(0.1*$(number of peak regions), 10). Motifs with a score less than 1.0 were considered to have low quality, so we excluded them from our further analysis.

With the enrichment level of 825 motifs in each group, we can compare if there are any different enriched motifs between different groups.

3.2.3.6 Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables. We conducted a fisher's exact multiple test on motif's enrichment levels in peak regions of Myc_Group 1 and Myc_Group 2. Total length of non-repeat sequences in ChIP-binding regions (i.e. n_{2B}) can be described as total number of all possible starting points of a motif site. Number of motif sites in ChIP-binding regions (i.e. n_{1B}) can be described as number of observed starting points of a motif site in ChIP-binding regions. Our null hypothesis is that the relative enrichment levels of the same motif in Myc_Group 1 and Myc_Group 2 are the same. Alternative hypothesis is that the relative enrichment levels of the same motif in Myc_Group 1 and Myc_Group 2 are different. For each motif, our dataset is presented in Table 3.2. The probability of obtaining any such set of values is given by the hypergeometric distribution:

$$p = \frac{\binom{n_{2B}}{n_{1B}}\binom{n'_{2B}}{n'_{1B}}}{\binom{n_{2B}+n'_{2B}}{n_{1B}+n'_{1B}}}$$

We conducted a Fisher's exact test, two-tailed, in the R statistical computing environment to check if there were any statistically significant differences of motif's enrichment levels between Myc_Group 1 and Myc_Group 2. The significance level is set at 0.25. Bonferroni correction is needed in the multiple Fisher's exact tests.

	Number of observed starting points of a motif site in ChIP-binding regions	Number of points that are not observed to be starting points of a motif site in ChIP-binding regions	Number of all possible starting points of a motif site in ChIP-binding regions
Myc_Group 1	n_{1B}	$n_{2B} - n_{1B}$	n_{2B}
Myc_Group 2	n_{1B}	$n_{2B} - n_{1B}$	n_{2B}
	$n_{1B} + n_{1B}$	$n_{2B} + n_{2B} - n_{1B} - n_{1B}$	$n_{2B} + n_{2B}$

Table 3.2: Data in a 2 × 2 contingency table for Fisher's exact test for each motif

3.3 Results

3.3.1 Common binding regions

Signals from both Myc IP and IgG controls were normalized, and binding regions were detected and visualized using CisGenome [14]. Intersection of the two peak lists from SC antibody and Epit antibody identified 885 common Myc binding regions in both data sets. Among 2452 Myc binding regions, 1564 regions were identified only by Santa Cruz (SC) anti-N-terminal Myc antibody. Among 1957 Myc binding regions, 1058 regions were identified only by Epitomics (Epit) monoclonal anti-N-terminal Myc antibody. Only about 25% of the binding regions were discovered by both antibodies. Thus, we wondered whether the rest binding regions uncovered by only one antibody were antibody-specific or not.

3.3.2 There is linear increasing relationship between Epit antibody and SC antibody

With the average probe intensity of each sample, we can check if the average probe intensities were normalized. We plotted scatter plots on every two samples of IgG controls, every two samples of SC IP and two samples of Epit IP and fitted simple linear regression models. From the scatter plots between every two of the samples, we found that normalization has been done (Figure 3.1, 3.2). The average probe intensities of each sample were within the interval (6, 14). And each sample correlated with others very well, with multiple R-squared ranging from 0.51 to 0.76. The correlation between every two samples was very high, thus we knew that the average probe intensities were normalized.

To reduce probe effect, we compared the difference of average probe intensities between IPs and IgG controls for each antibody rather than just comparing the IPs of the two antibodies. We plotted a scatterplot based on the difference of average probe intensities in the log_2 scales between IPs and IgG controls for each antibody. From the scatter plot of average probe intensity between Myc-Epit $log_2(IP/IgG)$ and Myc-SC $log_2(IP/IgG)$ (Figure 3.3), we found that the two antibodies correlated with each other with correlation coefficient of 0.39. Most of the data points were centered around the linear regression line, while there were only a few regions that might be antibody-specific.

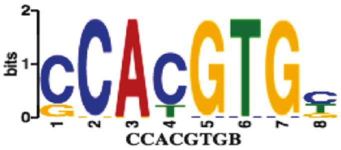
According to the output of the linear regression by limma, the minimum of adjusted p-values is 0.87 (Figure 3.5). All the adjusted p-values are larger than 0.05. From Student's t Q-Q plot (3.4), we found that the residuals were normally distributed. Thus, we concluded that there was no statistically significant difference of average probe intensities between the two antibodies. The two antibodies correlated with each other very well. As q-values were larger than 0.8, we cannot generate five groups according to the values of q-value. Because the p-value was uniformly distributed (Figure 3.5), we considered to perform the analysis and see if we could find some specific motifs with different preference for the two antibodies though not significant. Thus, we generated new five groups of peak regions based on p-values instead of q-values, with cutoffs same as described in methods. We performed de novo motif discovery analysis on five groups of peak regions and found enriched motifs in each group.

3.3.3 Enriched motifs in different groups and comparison of relative enrichment of enriched motifs.

After visual inspection of motif logos, redundant motifs were removed. Only one copy with the highest motif score of these redundant motifs was kept. By motif selection procedure of 300 motifs discovered by de novo motif discovery, the motif patterns of 13 enriched motifs were summarized and put in table 3.3. Myc motif was discovered from binding regions of each group. In Myc_Group 1, 8 motifs discovered by de novo motif discovery were enriched, Motif 2, Motif 29, Motif 39, Motif 52, Motif 125, Motif 131, Motif 133 (Myc) and Motif 247. In Myc_Group 2, 1 motif discovered by de novo motif discovery was enriched, Motif 260. Motif 209, which is a Myc motif, discovered by de novo motif discovery was enriched in Myc_Group 3. Motif 209, which is a Myc motif, discovered by de novo motif discovery was enriched in Myc_Group 4. In Myc_Group 5, 4 motifs discovered by de novo motif discovery were enriched, Motif 133 (Myc), Motif 256, Motif 270 and Motif 284.

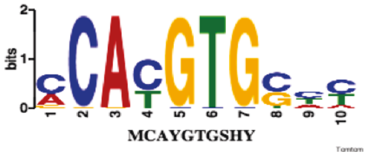
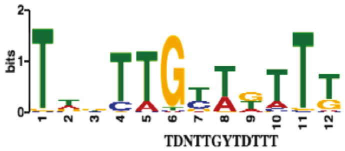
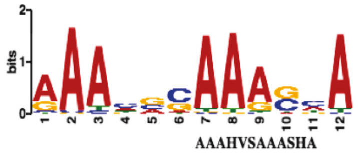
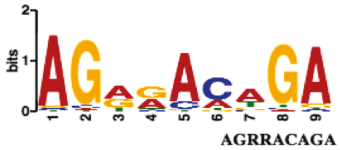
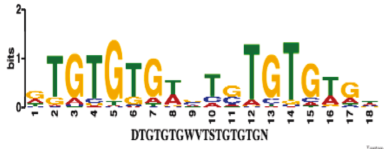
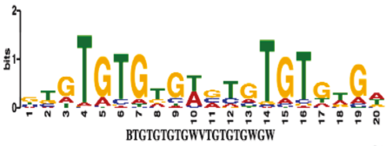
Then, with the relative enrichment levels of the 825 motifs, we first selected motifs enriched in Myc_Group 1 by requiring $r_1 \geq 2, n_{1B} \geq 10$ and compared the enrichment levels of these enriched motifs in five groups (Figure 3.6). Motif 2, Motif 29, Motif 39, Motif 52, Motif 125, Motif 131 and FOXO1 and PEA3 and IRF motifs were only enriched in Myc_Group 1. Next, we selected motifs enriched in Myc_Group 2 by requiring $r_1 \geq 2, n_{1B} \geq 13$ and compared the enrichment levels of these enriched motifs in five groups (Figure 3.7). Motif 260 was only enriched in Myc_Group 2. Then, we selected motifs enriched in Myc_Group 3 by requiring $r_1 \geq 2, n_{1B} \geq 162$ and compared the enrichment levels of these enriched motifs in five groups (Figure 3.8). The criteria for motifs enriched in Myc_Group 4 was $r_1 \geq 2, n_{1B} \geq 97$ simultaneously (Figure 3.9). USF, ARNT and EBOX motifs were found to be only enriched in Myc_Group 3 and Myc_Group 4. Finally, we selected motifs enriched in Myc_Group 5 by requiring $r_1 \geq 2, n_{1B} \geq 74$ simultaneously and compared the enrichment levels of these enriched motifs in five groups (Figure 3.10). Motif 256, Motif 270 and Motif 284 were enriched only in Myc_Group 5. Myc motif was enriched in binding regions of every group except for Myc_Group 2.

Table 3.3: Summary of enriched motifs discovered by De novo motif discovery, motif score, known motif, motif logos. Motif matrices are represented as sequence logos.

Motif id	Motif score	Known Motif	Motif sequence logos
Motif 209	5.47	MYC	

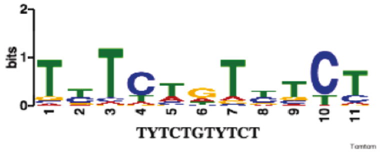
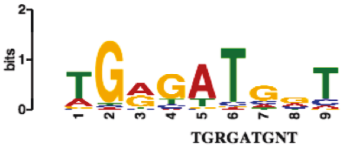
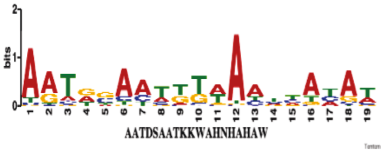
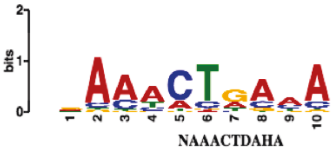
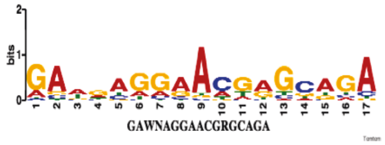
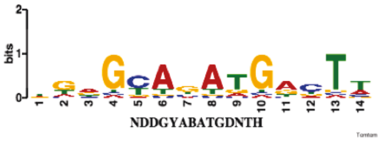
Continued on next page

Table 3.3 – *Continued from previous page*

Motif id	Motif score	Known Motif	Motif sequence logos
Motif 133	4.92	MYC	
Motif 131	3.53		
Motif 125	3.41		
Motif 260	2.86		
Motif 270	2.51		
Motif 284	2.22		

Continued on next page

Table 3.3 – *Continued from previous page*

Motif id	Motif score	Known Motif	Motif sequence logos
Motif 247	2.19		
Motif 39	1.93		
Motif 256	1.57		
Motif 29	1.47		
Motif 52	1.26		
Motif 2	1.09		

3.3.4 Differential DNA motifs between Myc_Group 1 and Myc_Group 2

As Myc_Group 1 and Myc_Group 2 both contain peak regions whose relative p-value of the linear model is less than 0.05, the only difference is the sign of t statistics. Regions in Myc_Group 1 have a higher binding intensity with SC antibody, while regions in Myc_Group 2 have a higher binding intensity with Epit antibody. We sought to find if there were any statistically significant differences of motif's enrichment level between Myc_Group 1 and Myc_Group 2.

After performing Fisher's exact test, we used Bonferroni family wise error rate to check differences of motif's enrichment level between Myc_Group 1 and Myc_Group 2. Having a total of 825 hypotheses, the Type I error with Bonferroni correction is 0.00030 (i.e. 0.25/825). Thus, by assuring $FWER \leq 0.25$, the probability of making even one type I error in the family is controlled at level 0.25.

There were two motifs that were of significantly different enrichment levels between Myc_Group 1 and Myc_Group 2, Motif 41 and Motif 52 (Table 3.4). They were all discovered from Myc_Group 1, and the enrichment levels of Motif 41 and Motif 52 were higher in Myc_Group 1. Our hypothesis is that Motif 41 and Motif 52 might be SC antibody specific binding motifs. Motif 41 seems not to be very informative. Thus, we will not discuss the function of Motif 41. We used TOMTOM motif comparison Tool to visualize the sequence logo of Motif 52, and noted that Motif 52 was very similar to

Sfpi1 motif (Figure 3.11). Sfpi1 is bound to Transcription factor PU.1. The function of Sfpi1 is binding to the PU-box, a purine-rich DNA sequence (5'- GAGGAA-3') that can act as a lymphoid-specific enhancer. Sfpi1, as a transcriptional activator, may be specifically involved in the differentiation or activation of macrophages or B-cells. As our ChIP-chip experiments were conducted in human B-cells, Sfpi1 motif might be involved in differentiation or activation of B-cells and specifically bound to Myc-SC antibody.

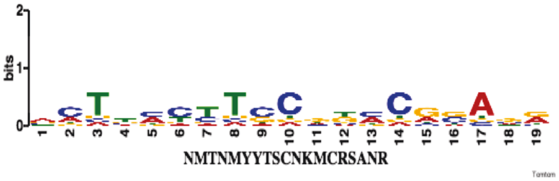
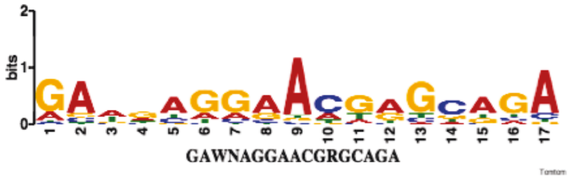
Motif id	Motif logo
Motif 41	
Motif 52	

Table 3.4: Motifs that were of significantly different enrichment levels between Myc_Group 1 and Myc_Group 2. Motif matrices are represented as sequence logos.

3.4 Summary and Discussion

3.4.1 Summary

By comparing the difference of average probe intensities between IPs and IgG controls for each antibody, we noted that there was linear increasing relationship between Epit and SC antibodies for MYC. After dividing the union of the two peak lists into five groups based on p-values, we found some motifs were enriched in individual groups. Motif 260 was only enriched in Myc_Group 2. Myc motif was enriched in binding regions of every group except for Myc_Group 2. The enrichment levels of Motif 41 and Motif 52 (Table 3.4) were found to be of significantly different between peak regions of Myc_Group 1 and Myc_Group 2. That is, Motif 41 and Motif 52 might be SC antibody-specific binding motifs. Motif 41 does not contain too much information. But Motif 52 is very similar to Sfpi1 motif, might be involved in differentiation or activation of B-cells and specifically bound to Myc-SC antibody.

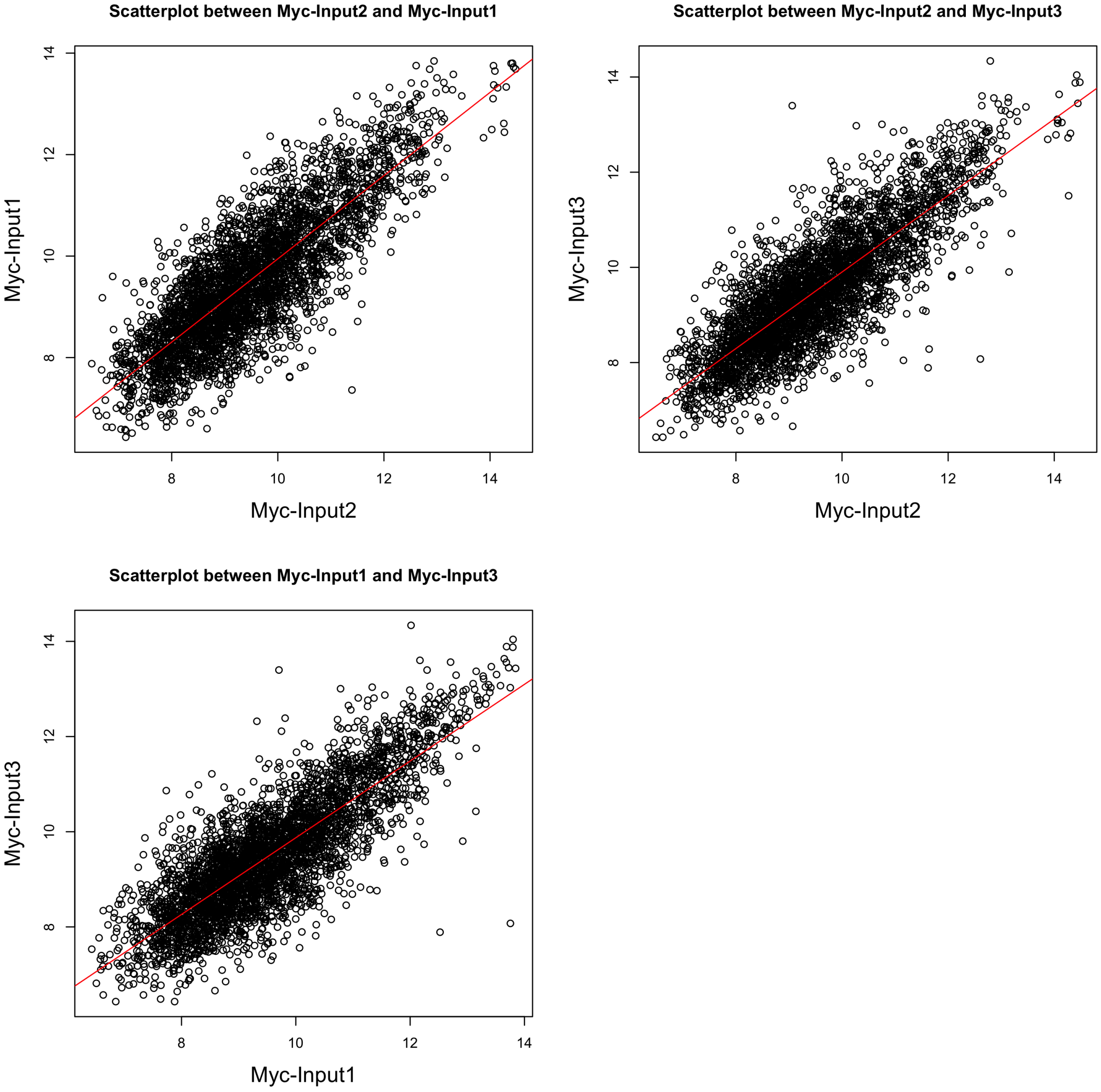
3.4.2 Discussion

By intersection of the two peak lists from SC and Epit antibody, we just identified 885 common Myc binding regions. At first, we considered that the two antibodies might not correlate with each other well. However, after fitting a linear regression of average probe intensities, we found that there was good correlation between Epit and

SC antibodies for MYC. Such a few common binding regions might be resulted from algorithm of detecting binding regions. According to previous studies, intersections of the two peak regions revealed by Epit and SC antibodies were used to determine binding targets of Myc [14]. As Epit and SC antibodies correlate with each other well, we could use union of the two peak regions to find binding targets of Myc. Analyzing based on intersections of two datasets might result in losing some binding targets of Myc. There are very few regions that are antibody specific. The interactions of Myc with different antibodies are distinct as they have different binding sites. The binding pattern of Myc with different antibodies can influence the binding of Myc to different DNA motifs in several ways. The binding of antibody with Myc might occupy the binding site of Myc with some particular DNA motifs. As a consequence, these regions cannot be immunoprecipitated by this antibody. The binding of antibody with Myc might give rise to deformation of Myc's structure, thus Myc could not recognized some specific DNA motifs any more. This kind of study can be beneficial to the choice of antibody when conducting a ChIP-chip experiment.

In addition, we found that Sfpi1 motif might be SC antibody-specific binding motif. Sfpi1, as a member of the Ets family, is expressed selectively on B cells, myeloid cells and macrophages [17]. PU.1/Sfpi1 regulates the expression of several genes, which is crucial for macrophage and B-cell differentiation. Sfpi1 plays an important roles in the renewal of progenitor cells and in early differentiation [18]. Published literature consistently refers to the regulation of PU.1 in the progenitors of blood cells

[18]. In our study, enrichment level of Sfpi1 motif is relatively higher in binding regions immunoprecipitated by SC antibody than in binding regions immunoprecipitated by Epit antibody for Myc. However, we are not sure if the significant difference of enrichment level of Sfpi1 motif is happened by chance or not. Further studies are needed to check if Sfpi1 motif is SC antibody-specific binding motif for Myc. Our findings suggest that future investigation of the motif is worthwhile. Furthermore, we should analyze other transcription factors to see if antibody-specific binding motifs really exist.



Scatterplot between Myc-Input2 and Myc-Input1

Scatterplot between Myc-Input2 and Myc-Input3

Scatterplot between Myc-Input1 and Myc-Input3

Figure 3.1: Scatter plot of every two replicates of Myc IgG controls

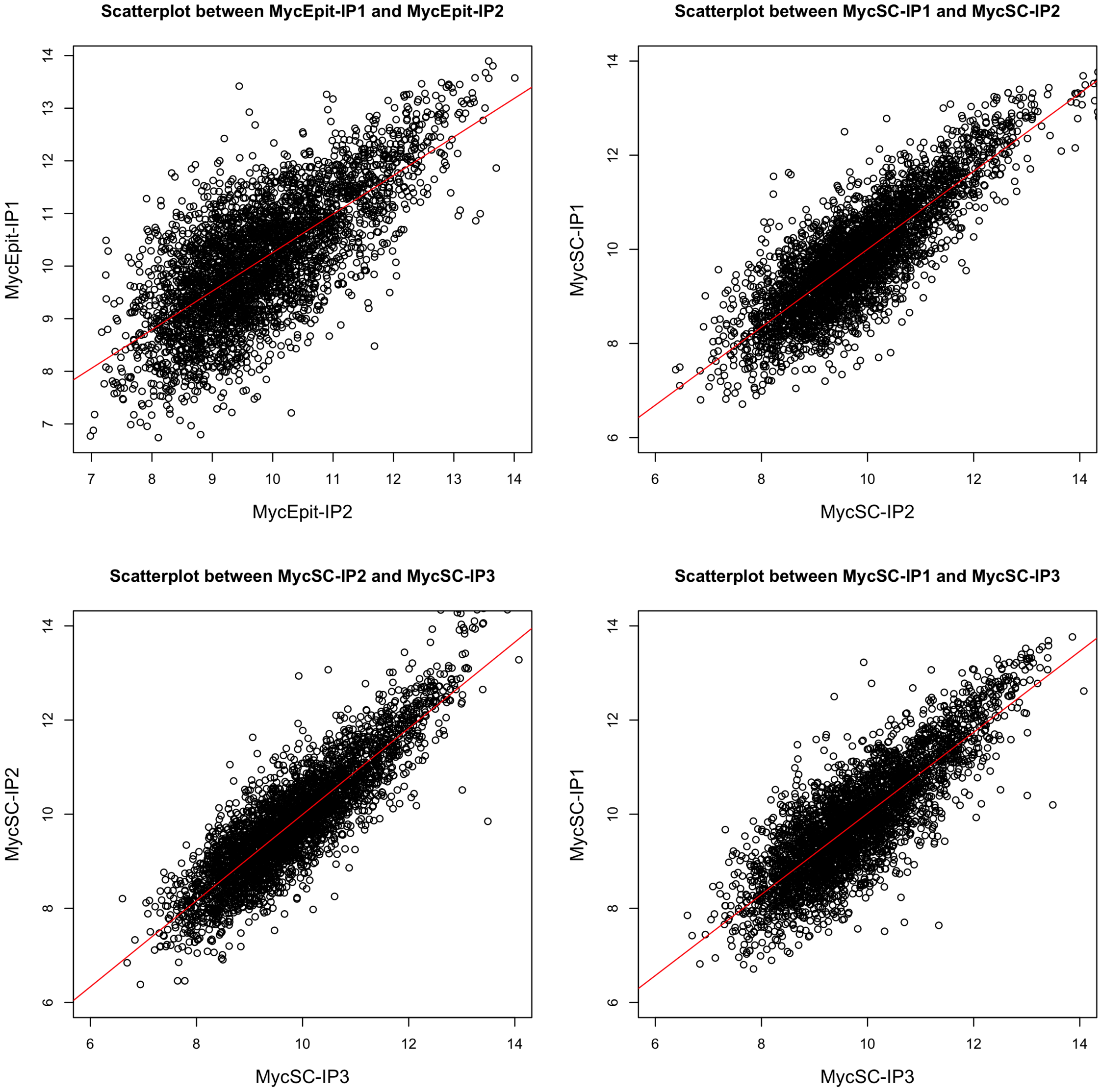


Figure 3.2: Scatter plot of every two replicates of MYC IPs.

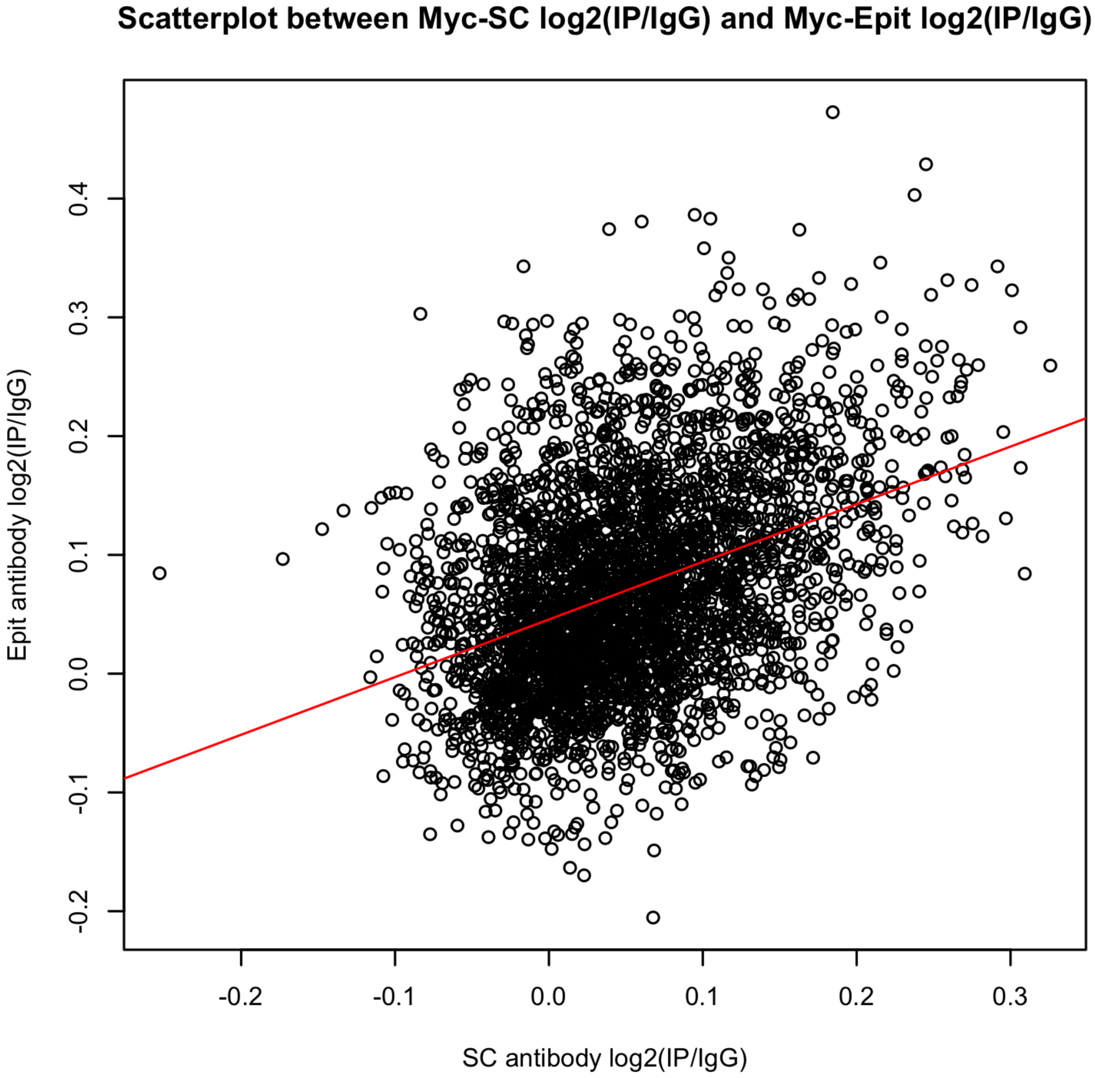


Figure 3.3: Scatter plot between Myc-Epit $log_2(IP/IgG)$ and Myc-SC $log_2(IP/IgG)$

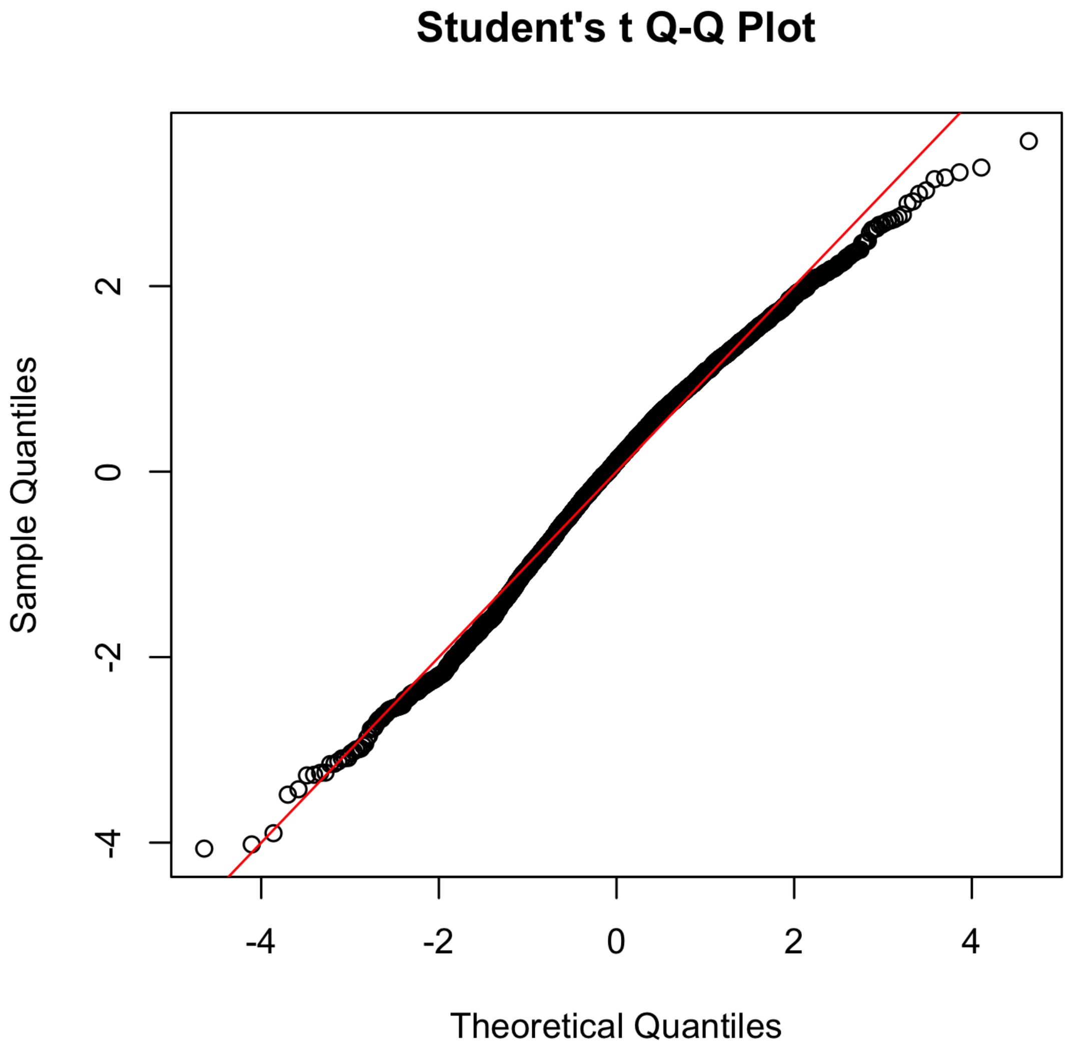


Figure 3.4: QQ plot of average probe intensities of IP samples.

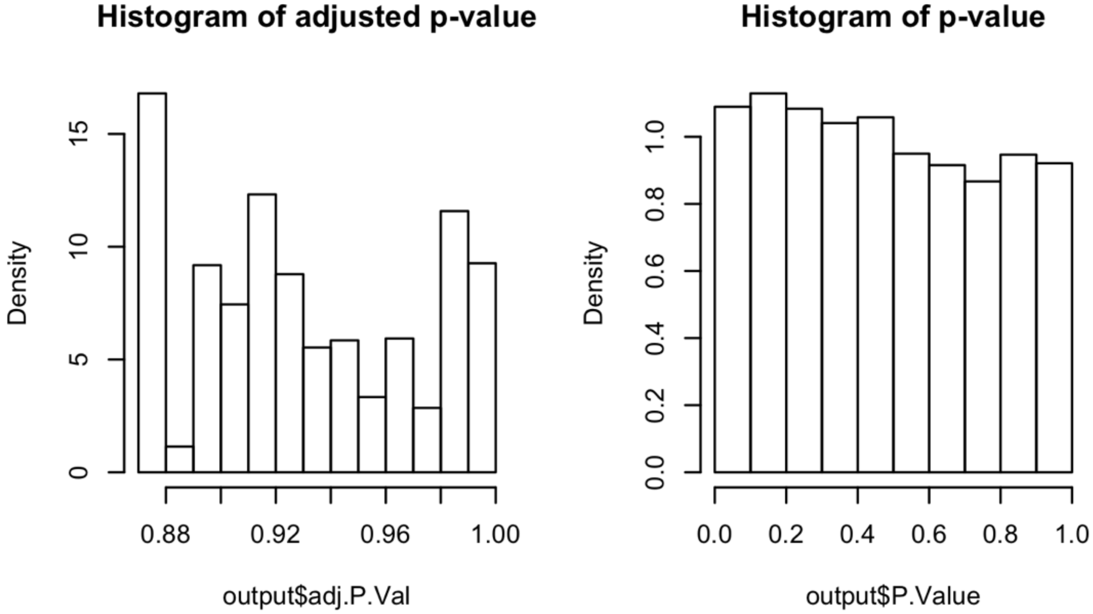


Figure 3.5: Histogram of adjusted p-value and p-value output from limma package in R

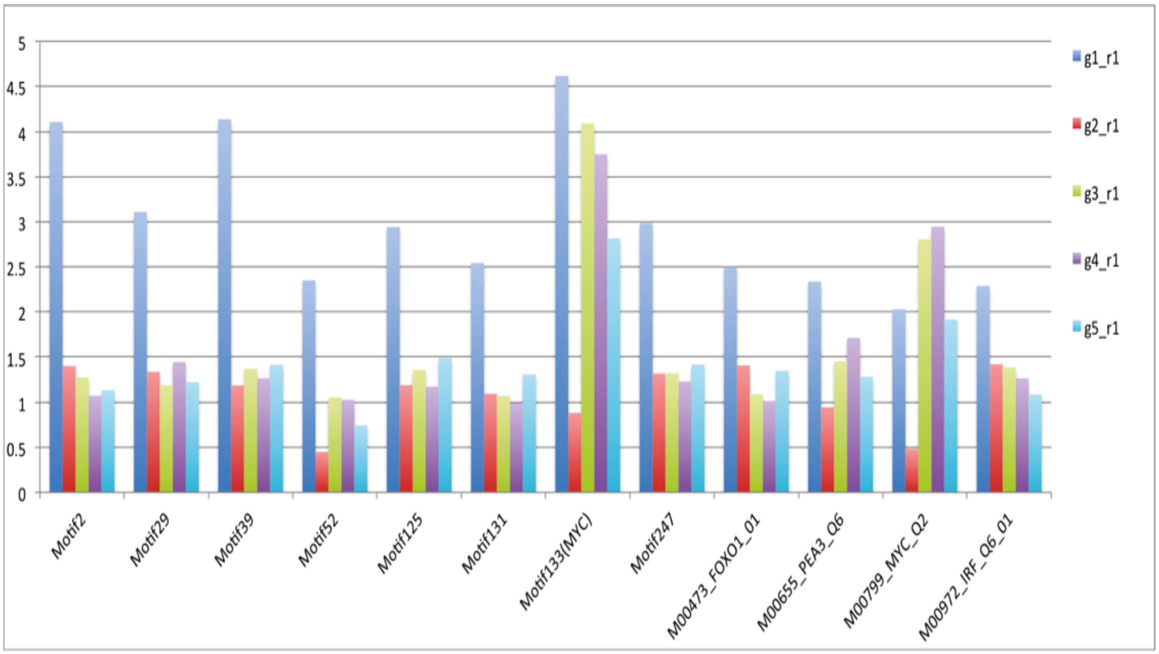


Figure 3.6: Comparison of relative enrichment level of enriched motifs in Myc_Group 1

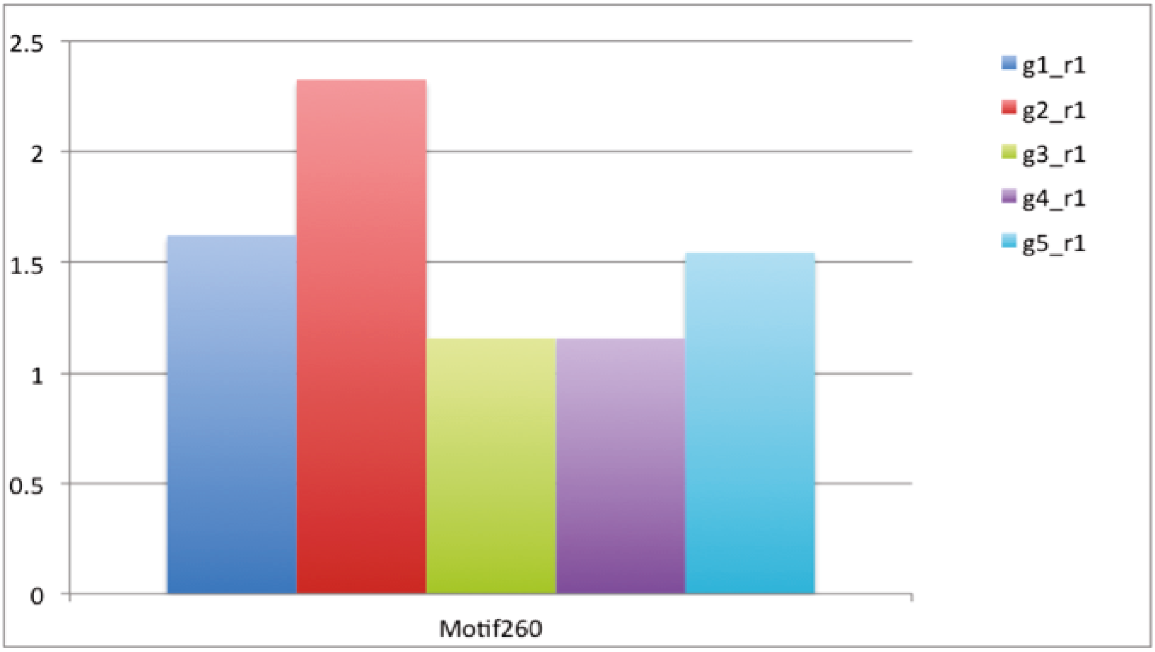


Figure 3.7: Comparison of relative enrichment level of enriched motifs in Myc_Group 2

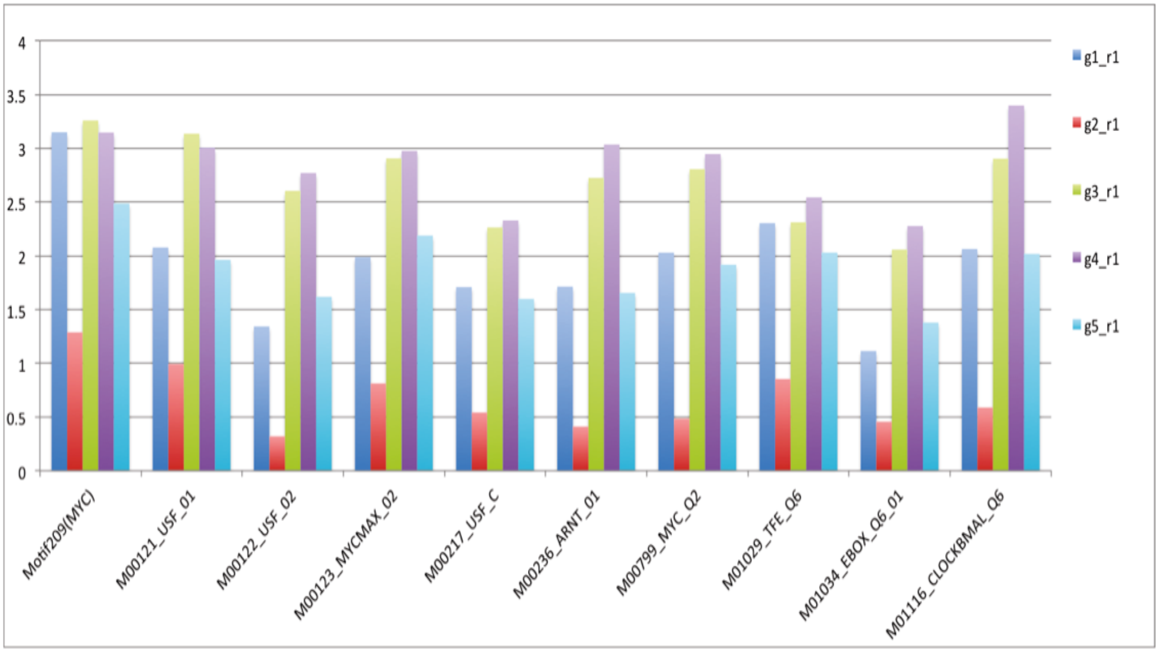


Figure 3.8: Comparison of relative enrichment level of enriched motifs in Myc_Group 3

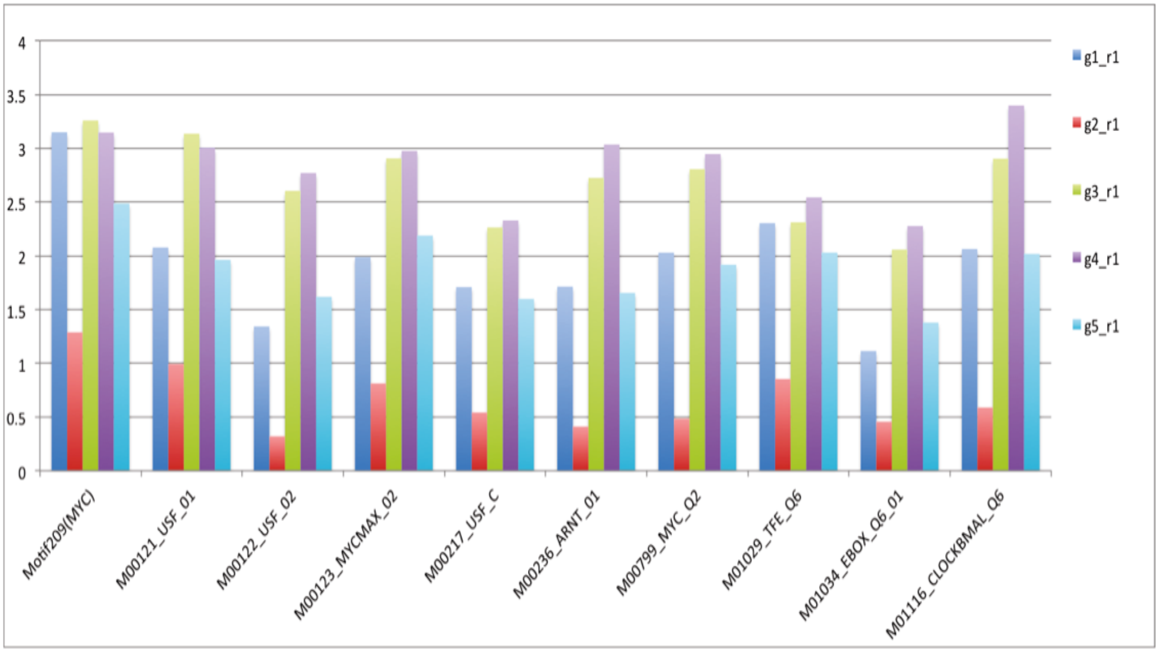


Figure 3.9: Comparison of relative enrichment level of enriched motifs in Myc_Group 4

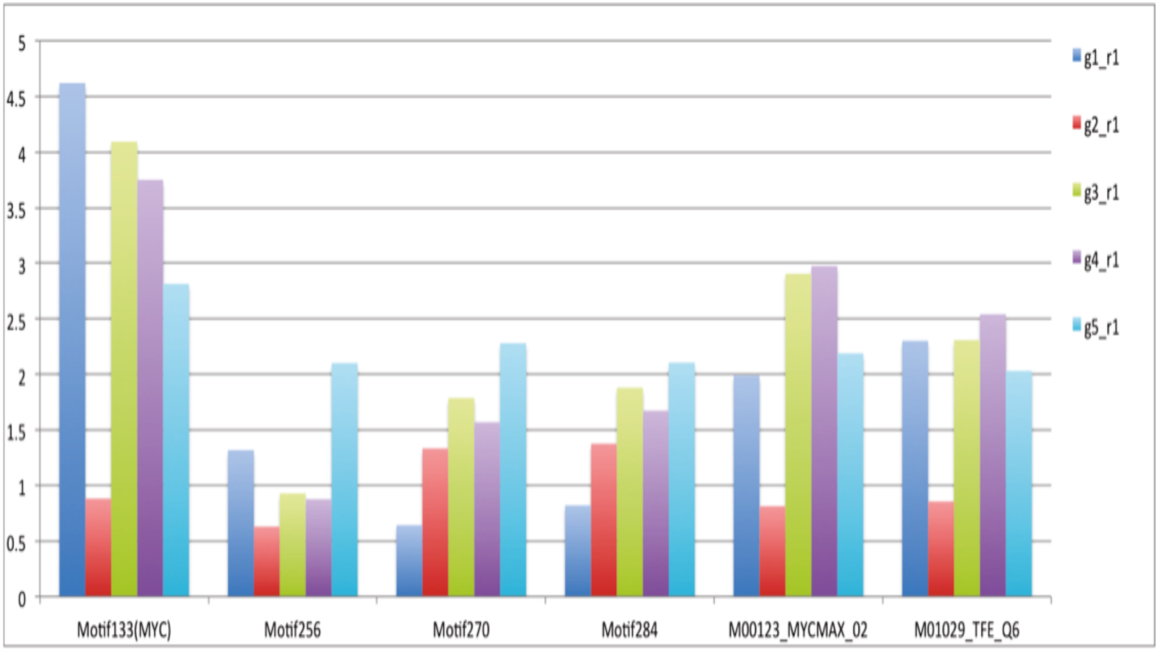


Figure 3.10: Comparison of relative enrichment level of enriched motifs in Myc_Group 5

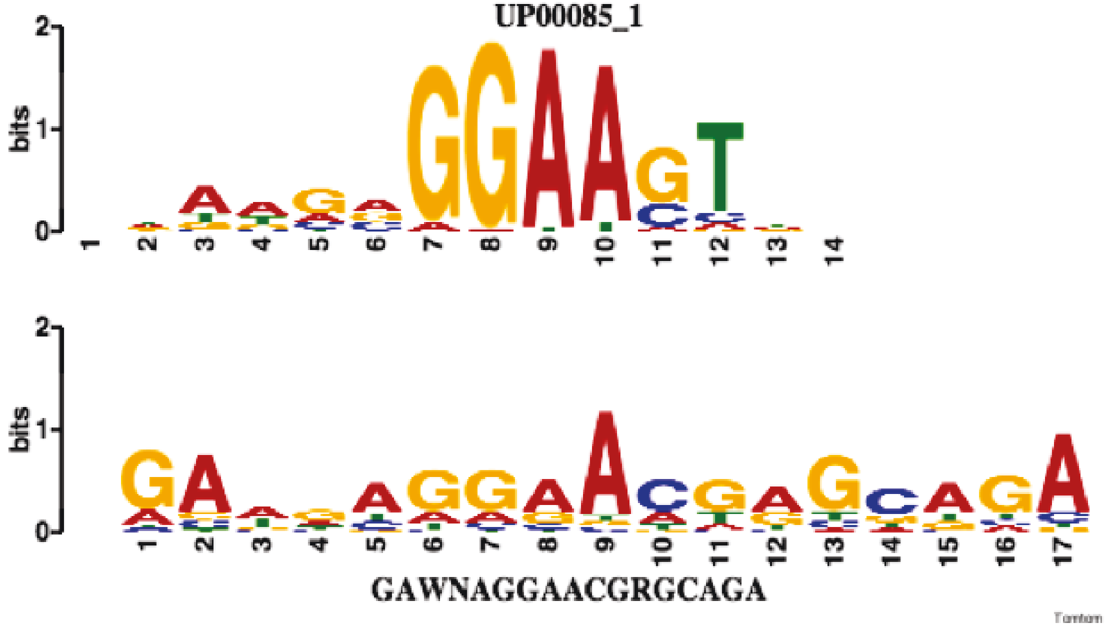


Figure 3.11: Comparison of motif logos: up is motif logo of Sfpi1, down is motif logo of Motif 52.

Bibliography

[1] S. Keleş, M. J. Van Der Laan, S. Dudoit, and S. E. Cawley, "Multiple Testing Methods For ChIP–Chip High Density Oligonucleotide Array Data," *Journal of Computational Biology*, vol. 13, no. 3, pp. 579–613, Apr. 2006.

[2] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, Nov. 2008.

[3] H. Ji and W. H. Wong, "Computational Biology: Toward Deciphering Gene Regulatory Information in Mammalian Genomes," *Biometrics*, vol. 62, no. 3, pp. 645–663, Aug. 2006.

[4] P. J. Park, "ChIP–seq: advantages and challengesof a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 669–680, Sep. 2009.

[5] R. Gottardo, W. Li, W. E. Johnson, and X. S. Liu, "A Flexible and Powerful Bayesian Hierarchical Model for ChIP–Chip Experiments," *Biometrics*, vol. 64, no. 2, pp. 468–478, 2008.

[6] H. Ji and W. H. Wong, “TileMap: create chromosomal map of tiling array hybridizations,” *Bioinformatics*, vol. 21, no. 18, pp. 3629–3636, Aug. 2005.

[7] J. S. Liu, Q. Zhou, X. S. Liu, and S. T. Jensen, “Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective,” *Statistical Science*, vol. 19, no. 1, pp. 188–204, Feb. 2004.

[8] S. A. Vokes, H. Ji, S. McCuine, T. Tenzen, S. Giles, S. Zhong, W. J. R. Longabaugh, E. H. Davidson, W. H. Wong, and A. P. McMahon, “Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning,” *Development*, vol. 134, no. 10, pp. 1977–1989, Apr. 2007.

[9] S. A. Vokes, H. Ji, W. H. Wong, and A. P. McMahon, “A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb.” *Genes & Development*, vol. 22, no. 19, pp. 2651–2663, Oct. 2008.

[10] H. Ji, S. A. Vokes, and W. H. Wong, “A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors.” *Nucleic Acids Research*, vol. 34, no. 21, pp. e146–e146, 2006.

[11] “Integrating regulatory motif discovery and genome-wide expression analysis.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3339–3344, Mar. 2003.

[12] B. Tarchini, D. Duboule, and M. Kmita, “Regulatory constraints in the evolution of the tetrapod limb anterior–posterior polarity,” *Nature*, vol. 443, no. 7114, pp. 985–988, Oct. 2006.

[13] V. Seitz, P. Butzhammer, B. Hirsch, J. Hecht, I. Gütgemann, A. Ehlers, D. Lenze, E. Oker, A. Sommerfeld, E. von der Wall, C. König, C. Zinser, R. Spang, and M. Hummel, “Deep Sequencing of MYC DNA-Binding Sites in Burkitt Lymphoma,” *PLoS ONE*, vol. 6, no. 11, p. e26837, Nov. 2011.

[14] H. Ji, G. Wu, X. Zhan, A. Nolan, C. Koh, A. De Marzo, H. M. Doan, J. Fan, C. Cheadle, M. Fallahi, J. L. Cleveland, C. V. Dang, and K. I. Zeller, “Cell-Type Independent MYC Target Genes Reveal a Primordial Signature Involved in Biomass Accumulation,” *PLoS ONE*, vol. 6, no. 10, p. e26057, Oct. 2011.

[15] D. Perna, G. Faga, A. Verrecchia, M. M. Gorski, I. Barozzi, V. Narang, J. Khng, K. C. Lim, W.-K. Sung, R. Sanges, E. Stupka, T. Oskarsson, A. Trumpp, C.-L. Wei, H. Müller, and B. Amati, “Genome-wide mapping of Myc binding and gene regulation in serum-stimulated fibroblasts,” *Oncogene*, vol. 31, no. 13, pp. 1695–1709, Aug. 2011.

[16] G. K. Smyth, “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments,” *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, Feb. 2004.

[17] J. Lloberas, A. Celada, and C. Soler, “The key role of PU.1/SPI-1 in B cells,

myeloid cells and macrophages," *Immunology Today*, vol. 20, no. 4, pp. 184–189, Apr. 1999.

[18] V. Olive, N. Wagner, S. Chan, P. Kastner, C. Vannetti, F. Cuzin, and M. Rassoulzadegan, "PU.1 (Sfpi1), a pleiotropic regulator expressed from the first embryonic stages with a crucial function in germinal progenitors." *Development*, vol. 134, no. 21, pp. 3815–3825, Nov. 2007.

Vita

Shilu Zhang received the degree of Bachelor of Science in Biological Sciences from Nanjing University in 2012. During her undergraduate education, she won First-class the People's Scholarship at Nanjing University in 2009. Her undergraduate's research focused on identifying the role of XAK1 gene in XA21-mediated innate immunity, and her thesis was honored as the Outstanding Bachelor's Thesis at Nanjing University in 2012. She was accepted into the ScM program offered by the Department of Biostatistics at Johns Hopkins Bloomberg School of Public Health in 2012. She won the department's Kocherlakota Award for best performance in the first-year comprehensive examination at Johns Hopkins Bloomberg School of Public Health in 2013, and received the 75% tuition scholarship in 2013. She is currently working on next-generation sequencing data analysis at Johns Hopkins Bloomberg School of Public Health.