INTEGRATION OF HETEROGENEOUS DATASETS FOR THE PREDICTION OF DIRECTLY REGULATED GENES

DENG NIANTAO

(B.Sc., Shanghai Jiao Tong University)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgements

I would like to take this opportunity to thank my supervisors, Associate Professor Choi Kwok Pui and Dr. Guillaume Bourque. Their advice and patience in the past one year are greatly appreciated. Thanks Guillaume for providing the data for our analysis.

My thanks also go to the Department of Statistics and Applied Probability, which provides us a wonderful place to study.

I would also like to thank Wei Xing and Gerald Wang, for their help and time in explaining the experiments, biological concepts and reviewing the thesis draft.

Finally, I wish to thank my family and friends for their support through the whole project.

Deng Niantao

August 2008

Contents

Summary

Transcription factors (TF) play critical roles in the system that controls transfer of genetic information from DNA to RNA. Estrogen Receptor α (ERα), which is the master transcriptional regulator of breast cancer phenotype, is of particular interest in understanding carcinogenesis of breast cancer. Some relevant biological concepts are introduced in Chapter 1.

In the process of transcription, transcription factors bind to DNA and regulate the gene expression. Various kinds of experiments have been devised to understand the mechanism of regulation. On one hand, experiments such as ChIP-ChIP and ChIP-PET analysis could be performed to map ERα binding sites on a whole genome scale, and consequently a group of high confidence binding regions could be identified. On the other hand, DNA microarray experiments can measure the level of expression for thousands of genes at the same time. In Chapter 2, we mainly describe four datasets studied in this thesis, including two groups of high confidence binding regions and two microarray gene expression profiles. We introduce the datasets separately for binding data and gene expression. For binding data, we explain an important concept that is used to measure binding strength and conduct

some preliminary analysis. A further analysis of concentration of the binding data will be introduced later in Chapter 3. As for gene expression data, besides the data description, we also introduce methods on gene selection, such as Welch t-test and Significant Analysis of Microarray (SAM). After that, we obtain a particular group of differentially expressed genes by SAM for our future analysis. Lastly the use of the UCSC database is also mentioned in this chapter.

The main concern in this thesis is to explore the association of these high confidence binding regions with gene expression data. In Chapter 3, our objective is to identify the rules that link transcription factor binding to the regulation of genes. The preliminary analysis shows the distribution of binding strength. In order to identify the impact of binding strength on the regulation of genes, we map the position of binding sites to the $5'$ and $3'$ end of regulated genes. We then obtain the occurrence of high confidence binding regions in the vicinity of the selected genes. By comparing the binding strength of binding sites in the neighborhood of regulated genes with that of all the binding sites, we show there is a positive impact of binding strength on the gene regulation. After that, we investigate the density of binding sites along the genome, using various lengths of windows to study the concentration of binding clusters. Similarly, we analyze the effect of the concentration on the gene regulation. Finally, we integrate all the possible factors impacted on gene regulation into a score function. And the accuracy of score function in separating expressed and control genes is evaluated by the ROC curve analysis.

In the last Chapter, we sum up the important conclusions in this thesis. And refer to our original question, we point out the limitation in our study and propose several ways for improvement. Also, a discussion of problems that encountered during the analysis and possible areas for future study will be highlighted.

List of Tables

List of Figures

Chapter 1

Biological Background

In this chapter we introduce some concepts of central importance, such as transcription factors, Estrogen Receptor, and relevant experiments for our datasets: Microarray experiment (for gene expression data) and Chromatin Immunoprecipitation (ChIP) (for data of binding sites).

1.1 Transcription Factor

The process of transcription in molecular biology refer to the synthesis of RNA from a particular segment of DNA through the function of RNA polymerase. A Transcription Factor (TF) is a protein which is involved in the transcription of genes. They usually bind to the part of DNA which controls the level of gene expression. The place on cellular DNA to which transcription factor can bind is called Binding Sites (BS). Typically, BS might be found in the vicinity of genes, and would be involved in activating transcription of genes (promoter elements), in enhancing the transcription of genes (enhancer elements), or in reducing the transcription of genes (silencers).

1

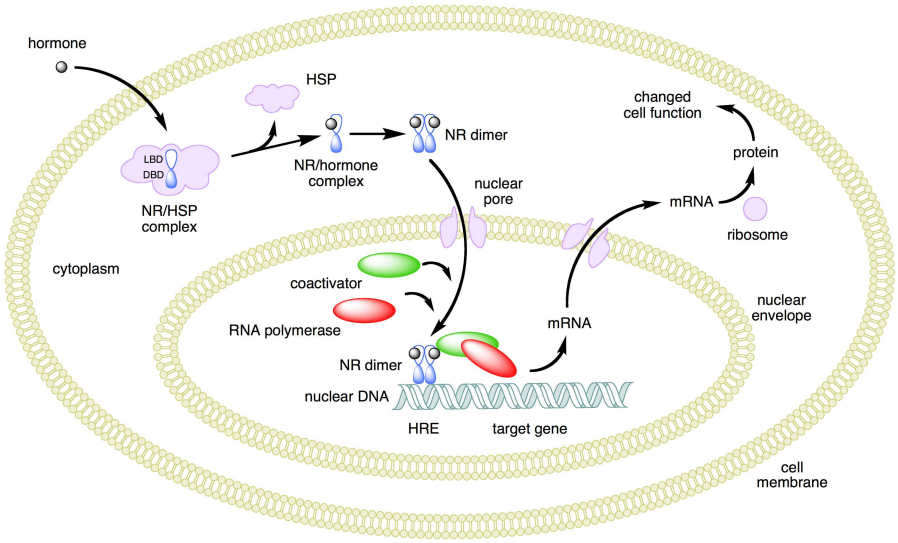

Figure 1.1: Mechanism of Nuclear Receptor Action

1.2 Estrogen Receptor α

Estrogen Receptors (ERs) (specifically ERα and ERβ) are ligand-dependent transcription factors that mediate cellular responses to estrogen (such as estradiol) in vertebrate development, physiological processes, and endocrine-related diseases.

[1] The figure depicts the mechanism of a class I nuclear receptor (NR) which, in the absence of ligand, is located in the cytosol. Hormone binding to the NR triggers dissociation of heat shock proteins (HSP), dimerization, and translocation to the nucleus where the NR binds to a specific sequence of DNA known as a hormone response element (HRE). The nuclear receptor DNA complex in turn recruits other proteins that are responsible for transcription of downstream DNA into mRNA which is eventually translated into protein which results in a change in cell function.

ERα, in particular, has been implicated in the etiology of breast cancer and is a major prognostic marker and therapeutic target in disease management. In general, ER is a kind of nuclear receptor, Figure 1.1 [2]shows the mechanism of NR action.

1.3 Microarray Experiment

Microarrays are widely used to measure gene expression differences across samples. They are able to study the expression patterns of thousands of genes and the interaction among the genes when they are put under the same experimental environment. There are two kinds of gene expression data. It can be either sequencing or hybridization based. Sequencing-based approaches include sequencing of complementary DNA (cDNA) libraries and serial analysis of gene expression (SAGE). While hybridization-based methods, such as Southern and Northern blots, colony hybridization, and dots blots, have long been used to identify and quantify nucleic acids in biological samples [Lee, 2004].

Analysis tools

Affymetrix analysis software is used to perform the preliminary probe-level quantitation of the microarray data. These data are further normalized using the RMA [Irizarry et al., 2003] normalization method.

Time course data

From the time course microarray expression data, differentially expressed genes are identified at each time point separately using the three untreated samples at the time point as controls against the three treated samples. The SAM [Parmigiani et al., 2003] statistical method is used to select differentially expressed genes. Genes are selected

[2]http://en.wikipedia.org/wiki/Image:Nuclear_receptor_action.png on July 9, 2008

based on a q-value with a specified cutoff.

1.4 ChIP Experiment

Chromatin Immunoprecipitation (ChIP) is a method for isolating and characterizing the specific pieces of DNA out of an entire genome, to which a protein of interest is bound. There are two common ways to characterize the DNA isolated: ChIP-ChIP and ChIP-Sequencing.

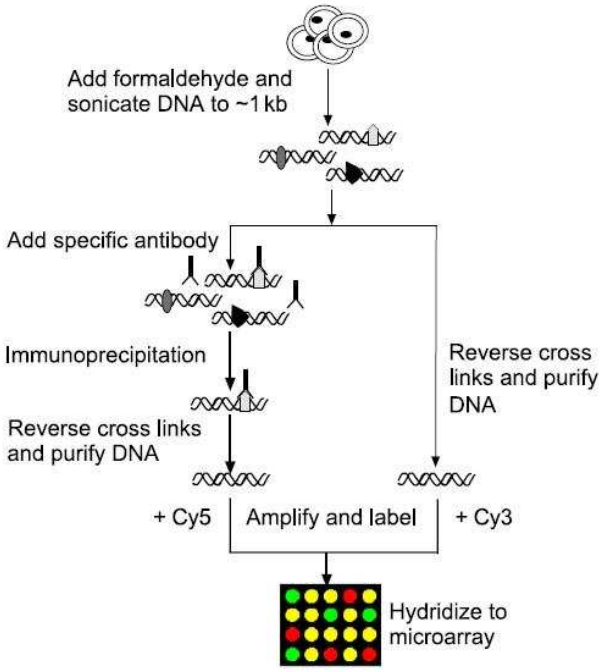

Figure 1.2: Summary of the ChIP-ChIP Procedure[Buck and Lieb, 2004]

1.4.1 ChIP-ChIP

In this variant, the DNA isolated from a ChIP experiment is characterized by labeling it with a fluorescent dye, then hybridizing it to a DNA array. Array

spots that "light up" are taken as evidence that their specific sequence is present in the ChIP product. Figure 1.2 shows the procedure of ChIP-ChIP experiment. We notice that enriched DNA from IP with protein-specific antibodies and DNA fragments direct from IP input are labeled by two different colors of fluorescent molecules (Cy5 and Cy3), after that they are combined and hybridized into a single DNA microarray chip. To design these arrays requires that one need to have some idea of what to expect in the ChIP isolated DNA.

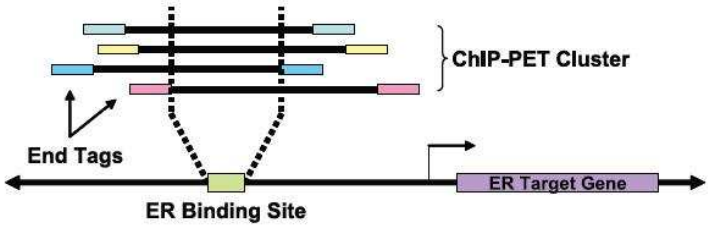

Figure 1.3: The Maximum Overlap PET [Lin et al., 2007]

1.4.2 ChIP-Sequencing

Under this variant, one can simply sequence every DNA fragments that immunoprecipitated with the antibody. An related sophisticated technology known as ChIP Pair End-Tagging (ChIP-PET) [Wei et al., 2006], characterizes unique DNA fragments and establish overlapping PET clusters to select high confidence binding sites clusters. Our datasets for ER binding sites (in Chapter 2) are obtained by ChIP-PET technology, which is targeted to map ERα binding sites in MCF-7

human breast cancer cells. An important concept of the experimental result is maximum overlap PET number (MoPET). The ChIP-PET experiment identifies groups of potential binding sites, which are in the unit of binding cluster. In each unit, the potential sites are overlapped with each other, the maximum overlapped region of all the sites define the start and termination position of this cluster. (for instance in Figure 1.3, the number of MoPET is 4.)

Chapter 2

Data Description

2.1 Binding Sites Data

Binding sites are places on the DNA to which a protein (such as transcription factor) can bind. ChIP-PET Analysis has been applied to map ER binding sites across the whole genome. Hormone-deprived MCF-7 cells were treated with 10nM estradiol for 45 minutes, and then DNA-bound receptor complexes were isolated through ChIP using anti-ERα antibodies [Lin et al., 2007].

After the quality of ChIP DNA fragments has been verified, the PET library was generated. The distinct PET Clusters were selected and a group of high confidence binding sites clusters were identified. All the ERα binding regions are located in every chromosome in the human genome, except for the Y chromosome, which is not present in MCF-7 cells from a female breast cancer patient.

2.1.1 ChIP-PET Data

There are two datasets for binding sites, both obtained by the ChIP-PET experiment. The first dataset of ER binding sites (Data I for short) was obtained

from [Lin et al., 2007]. It contains 1234 high confidence binding sites clusters, each binding cluster has a start, middle, end position and a maximum overlap PET (MoPET, definition refers to Chapter 1) size. The high confidence binding sites clusters have a high degree of overlapping, and for each cluster the MoPET ranges from 3 to 107.

Compared to the first dataset, the second one (unpublished) (Data II for short) is more precisely sequenced and is fixed with a cluster length of 200bp. It contains as many as 21,047 binding clusters. The data has the form:

Cluster ID	Chromo	Start	End	Middle	Mo-PET
714871	chr1	715036	715236	715136	11
5649376	chr1	5650153	5650053	5650253	11
...	...	...	...	...	...

where each Binding Cluster contains a group of Binding Sites identified by ChIP-PET experiment. "Start" is the start position of the overlapped region, and "End" stands for the termination position for the overlapped region. MoPET value in this dataset ranges from 8 to 228. Table 2.1 summaries the basic information of the two binding data, including cluster length and between clusters distance.

Distance between clusters						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Data I	530	206400	965600	2292000	2828000	37710000
Data II	513	3616	13320	138700	76700	28620000

Table 2.1: Five number summary for Between Cluster Distance

2.1.2 Preliminary Analysis of Binding Sites Data

Tables 2.2 and 2.3 show the distribution of Maximum Overlap number in each PET cluster(MoPET) for Data I and II respectively.

MoPET No.	Counts	Percentage	MoPET No.	Counts	Percentage
3	552	0.447	12	11	0.009
4	245	0.199	13	8	0.006
5	134	0.109	14	5	0.004
6	95	0.077	15	6	0.005
7	66	0.053	16	1	0.001
8	38	0.031	17	4	0.003
9	24	0.019	18	2	0.002
10	26	0.021	>18	9	0.007
11	8	0.006	Total	1234	1.00

Table 2.2: MoPET Distribution for Data I

From the tables, we can see both of the low MoPETs in the two datasets constitute the majority of all the binding clusters. Because of the large number of binding sites with low MoPET values- which may mean less significant binding sties, we would like to start with higher quality and stronger binding sites for our further analysis. And since the Data I contains only 1234 binding clusters (even less after removing low MoPET), we will later use only Data II to analyze the association between binding strength and gene regulation in Chapter 3. Thus, by choosing a cutoff of ≥ 11 for data II, we obtain 4870 binding clusters.

MoPET No.	Counts	Percentage	MoPET No.	Counts	Percentage
8	10049	0.477	18	159	0.008
9	4026	0.191	19	128	0.006
10	1922	0.091	20	117	0.006
11	1076	0.051	21	102	0.005
12	655	0.031	22	111	0.005
13	477	0.023	23	81	0.004
14	364	0.017	24	60	0.003
15	258	0.012	25	81	0.004
16	266	0.013	> 25	748	0.036
17	187	0.009	Total	21047	1.00

Table 2.3: MoPET Distribution for Data II

2.2 Identification of ER regulated genes

2.2.1 Introduction

Microarray can measure the expression of thousands of genes to identify changes in expression between different biological states. Methods are needed to determine the significance of these changes. In this chapter we will apply Welch t-test and Significance Analysis of Microarray (SAM) [Parmigiani et al., 2003] to select differentially expressed genes. To select the differentially expressed genes is important because only through those genes can we identify the mechanism of transcription. In order to explore more in-depth information of the expression data, normalization of the data is necessary to remove the "noise". There are several ways to normalize the data, and our data is normalized by the Robust Multiarray Average (RMA) [Irizarry et al., 2003] method. Using the normalized data, we apply the

SAM method and select differentially expressed by choosing a cutoff for False Discovery Rate (FDR). The genes selected will be used as potential regulated genes for further analysis. The discussion of association of binding sites with these potential regulated genes will be introduced in the next chapter.

2.2.2 Gene Expression Data

We include two gene expression datasets in our analysis. The first human gene expression data were obtained from the collection of ER in the whole human genome ($BrownLabDatasets$)[1].

It contains 23,597 gene expression profiles by microarray analyses, which are performed in triplicate over an estrogen stimulation time course ($0, 3, 6$ and $12h$), with 3h representing immediate transcription targets and both 6 and 12 representing delayed targets. Figure 2.1 shows the distribution of early expression data at 3h point. The expression data are analyzed using the RMA algorithm with the newest probe mapping, and the Welch t statistic is used to calculate the level of differential expression at each time point relative to 0 h [Carroll et al., 2006].

The second gene expression is from Genome Institute of Singapore [Lin et al., 2007] with a number of 54,675 probesets. This time course experiment contains three replications for both treated and untreated samples at 12h, 24h, 48h time points (details of the data in .CEL file is available) [2]. It is also normalized by RMA method (with background correction, quantile normalization, and log transformation). Figure 2.1 shows the distribution of early expression data (3h for Carroll's and 12h for Lin's) of both datasets.

[1] $http://research.dfci.harvard.edu/brownlab/datasets/index.php?dir=ER\_whole\_human\_genome/$

[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11352

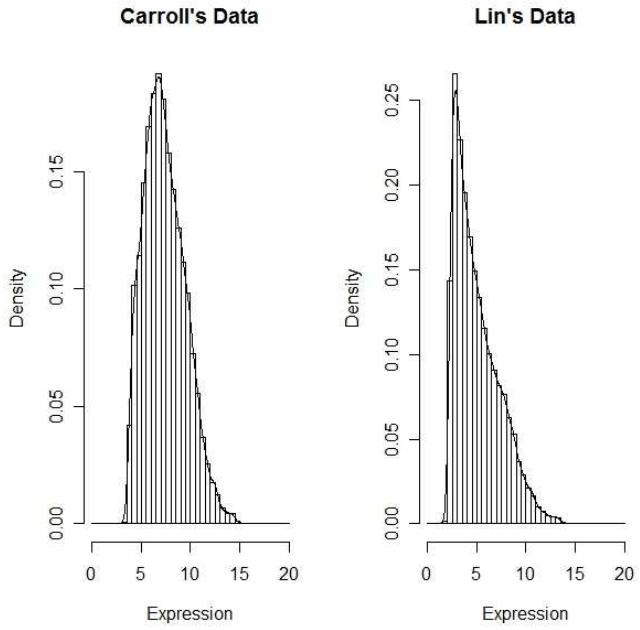

Figure 2.1: Early Expression Data

2.2.3 Significance Analysis of Microarray (SAM)

Methods based on conventional t tests provide the probability that a difference in gene expression occurred by chance. Although $p = 0.01$ is significant in the context of experiments designed to evaluate small number of genes, a microarray experiment for 10,000 genes would identify 100 genes by chance. This problem signals to a necessity to find some method specially designed for microarray analysis.

SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific t tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with score greater than a threshold are deemed potentially significant. The percentage of such genes identified by chance is the FDR, which is defined as:

Definition 1. FDR = $\mathbf{E}[V/R|R > 0]\mathbf{Pr}(R > 0)$

where V is the number of Type I error (false positives), S is the number of true positives, R = V + S is the total number of significant hypotheses (total positives).

2.2.4 Modified T-Test

Suppose that there are J genes measured on I arrays under two different experimental conditions. Let $\bar{x}_{j1}$ and $\bar{x}_{j2}$ be the average gene expression for gene j under condition 1 and 2, and let s_j be the pooled standard deviation for gene j:

$$s_j = \sqrt{(\frac{1}{I_1} + \frac{1}{I_2}) \cdot \frac{\sum_1 (x_{ji} - \bar{x}_{j1})^2 + \sum_2 (x_{ji} - \bar{x}_{j2})^2}{I - 2}}$$

Here, I_k is the number of arrays in condition k, and each summation is taken over its respective group. Then, a reasonable test statistic for assessing differential gene expression is the standard (unpaired) t-statistic:

$$t_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j}.$$

However, at low expression levels, the test statistic can be high because of small values of s_j, and consequently raises the false positive rate. We introduce a modified statistic to solve this problem:

$$d_j := \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s_0}.$$

the coefficient of variation of d_j was computed as a function of s_0 across the data and s_0 is chosen to minimize the coefficient of variation [Tusher et al., 2001]. The modified t-test would ensure that variance of d_j is independent of gene expression and also it would dampen large values of d_j that arise from low gene expression levels.

The SAM Procedure

1. Compute the ordered statistics

$$d_{(1)} \leq d_{(2)} \cdot \cdot \leq d_{(J)}.$$

2. Take B permutations of the group labels. For each permutation b ($1 \leq b \leq B$) compute statistics ${d_j}^{*b}$ and the corresponding order statistics

$${d_{(1)}}^{*b} \leq {d_{(2)}}^{*b} \cdot \cdot \leq {d_{(J)}}^{*b}.$$

   From the set of B permutations, estimate the expected order statistics by

$$\bar{d}^*_{(j)} = \frac{1}{B}\sum_{b=1}^{B} {d_j}^{*b}$$

   for $j = 1, 2, ...J$.

3. Plot the $d_{(j)}$ values versus the $\bar{d}^*_{(j)}$. For a fixed threshold Δ, starting at the origin, and moving up to the right, find the first $j = j_2$ such that

$$d_{(j)} - \bar{d}_{(j)} \geq \Delta.$$

   All genes past j_2 are called "significant positives". Similarly, start at the origin, move down to the left and find the first $j = j_1$ such that

$$d_{(j)} - \bar{d}_{(j)} \leq \Delta.$$

   All genes past j_1 are called "significant negatives". For each Δ, define the upper cut point $t_2(\Delta)$ as the smallest d_j among the significant positive genes, and similarly define the lower cut point $t_1(\Delta)$.

The figure shows an example of SAM selection by Stanford Tools [3]. The green points on the left below the cutoff and red points above the cutoff on the right stands for negative and positive regulated genes respectively.

[3] http://www-stat.stanford.edu/∼tibs/SAM/

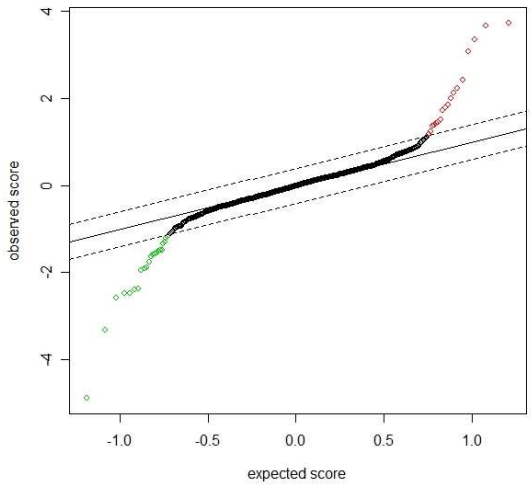

Figure 2.2: Significance Analysis of Microarray: an example plot from stanford SAM tools

2.2.5 Estimation of FDR and the *q-value*

Estimation of FDR

For a fixed rejection (fixed Δ), the FDR and pFDR are

$$FDR(\Delta) = \mathbf{E}[\frac{V(\Delta)}{R(\Delta)} \mid R(\Delta) > 0]\mathbf{Pr}(R(\Delta) > 0)$$

$$pFDR(\Delta) = \mathbf{E}[\frac{V(\Delta)}{R(\Delta)} \mid R(\Delta) > 0]$$

where

$V(\Delta) = \sharp\{d_j : gene\ j\ unchanged\ and\ d_j \leq t_1(\Delta)\ or\ d_j \geq t_2(\Delta)\}$,

$R(\Delta) = \sharp\{d_j : d_j \leq t_1(\Delta)\ or\ d_j \geq t_2(\Delta)\}$.

[Storey, 2002] develops the following estimates of the FDR and pFDR for a given Δ:

$$\widehat{FDR}_{\Delta'}(\Delta) = \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{R(\Delta) \vee 1},$$

$$\widehat{pFDR}_{\Delta'}(\Delta) = \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{Pr(R^0(\Delta) > 0) \cdot [R(\Delta) \vee 1]},$$

where

$$R^0(\Delta) = \frac{\sum_{b=1}^{B} \sharp\{d_j^b : d_j^b \leq t_1(\Delta) or d_j^b \geq t_2(\Delta)\}}{B},$$

$$Pr(R^0(\Delta) > 0) = \frac{\sharp\{b : \sharp\{d_j^b : d_j^b \leq t_1(\Delta) or d_j^b \geq t_2(\Delta)\} > 0\}}{B}.$$

And $\widehat{\pi}_0(\Delta')$ is an estimate of the overall proportion of true null hypotheses (unchanged genes). This estimate depends on our choosing another Δ'. In SAM it takes Δ' such that $R^0(\Delta') = J/2$ (i.e., half the null statistics fall in the rejection region defined by Δ'). The estimate is defined as

$$\widehat{\pi}_0(\Delta') = \frac{J - R(\Delta')}{J - R^0(\Delta')}.$$

Estimation of the q-value

$$\hat{q}\text{-value}(\text{gene}j) = \min_{\{\Delta:\text{gene j significant}\}} \widehat{pFDR}_{\Delta'}(\Delta).$$

The q-value of a particular gene can be estimated by taking the minimum $\widehat{pFDR}_{\Delta'}(\Delta)$ over all Δ for which the gene is found to be significant. The q-value estimate is conservatively consistent under the condition that is assumed in [Storey, 2002]. In testing for differential gene expression, we estimate q-value for each gene and it gives us a measure of strength of evidence for differential gene expression in terms of pFDR. This is an individual measure for each gene that simultaneously takes into account the multiple comparison. Note that by using the

q-value, the delta is chose to reach the minimum value for $\widehat{pFDR}_{\Delta'}(\Delta)$ (among all Δ that make the gene identified as significant). Therefore, it is not necessary to pick the rejection region or the desired error rate beforehand [Parmigiani et al., 2003].

Selected Regulated Gene Data

From the original gene expression data stated in Chapter 2, different expressed genes were selected by SAM based on a q-value of 2% [Lin et al., 2007]. After removing redundancy, we got 649 unique up-regulated genes and 624 down-regulated genes. These genes are of high importance and will later be associated with the binding sites data.

2.3 UCSC KGs Database

The University of California Santa Cruz(UCSC) Known Gene (KG) database is used to find the transcription start sites and end sites of genes in the profile, as well as other useful information like geneID, strand, chromosome number, etc. To obtain relevant information on the interested genes, we can upload a list of gene identifiers to the genome browser[4] and choose the relevant fields which we need to use.

In our data analysis, we use the probe identifiers from Expression data to locate the corresponding genes in the UCSC KG database. When comparing the property of selected genes with background, we use KG database hg17 (May 2004), which contains 37,859 genes, as the background for simulation.

[4]http://genome.ucsc.edu/

Conversion of Regulated Genes

It should be noted that both Data I and the regulated gene data are stored under hg17 (May 2004), but Data II is stored under hg18 (March 2006). Thus, we need to convert the gene data to hg18 when associating Data II with regulated genes, by using the liftover tool under utilities in UCSC Genome Browser.

Chapter 3

Association of Binding Data with Gene Expression Data

3.1 Introduction

In this chapter, we aim to identify the rules that link TF binding sites to gene regulation. The association is explored by distance (distance to transcription starting sites (TSS)), binding strength (the MoPET value) and concentration of binding sites. To begin with, we map the position of binding clusters to the vicinity of regulated genes and analyze the distribution of their distances to TSS. Then we compare our result with [Lin et al., 2007] and give our observations. Moreover, we analyze the binding strength of those binding sites which are in the neighborhood of regulated genes' TSS. And we conclude that the binding clusters with a higher MoPET value are more prone to be associated with regulated genes. Finally, we come up with a scoring function for genes, which includes all the potential factors we identified in previous study. Simulation is conducted in the UCSC KGs database (hg17) to verify the scoring function: we score a random set of genes (of the same number as our potential regulated genes) and compare their scores with

potential regulated genes.

3.2 Association of Binding Sites with Gene Expression Data

3.2.1 Mapping to Regulated Genes

In order to associate the Binding Sites data with Gene Expression data, we mapped the location of the binding sites relative to the start and termination sites of E2 up- and down-regulated genes[1].

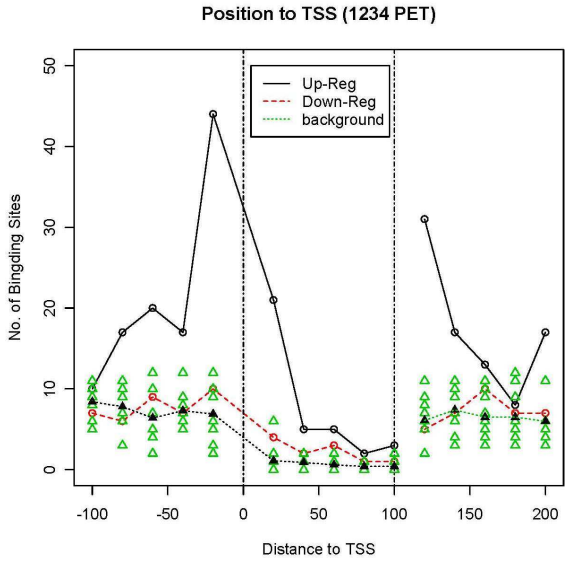

Figure 3.1: Position Relative to Transcription Starting Sites (1234 B.C.)

We mapped two binding sites data (Data I and Data II) to the 5' and 3' position of regulated genes respectively. The distances are measured in 20kb interval in

[1]The Gene Expression Data is from [Lin et al., 2007]

the region of 100kb upstream to 100kb downstream. Figure 3.1 is for Data I and Figure 3.2 is for Data II. As shown in Figure 3.1, approximately 45 ER binding clusters were found within 20kb of the transcriptional starting sites of up-regulated genes, while only 10 ER binding clusters were found for the down-regulated genes. The background was simulated for 700 randomly selected genes from UCSC KGs database, which used as a reference.

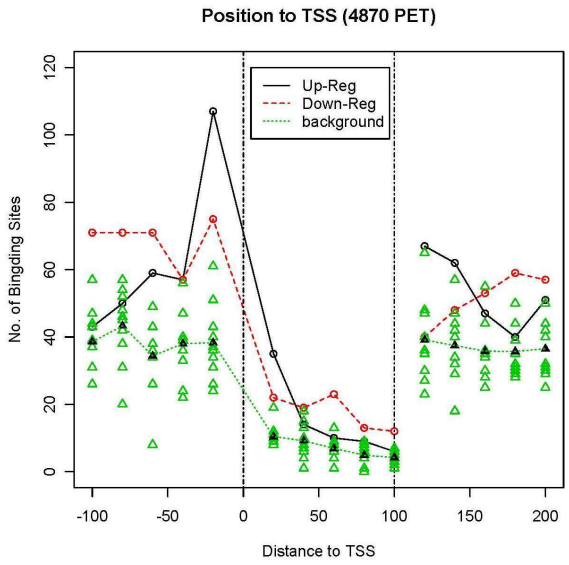

Figure 3.2: Position Relative to Transcription Starting Sites (4870 B.C.)

In Figure 3.2, the same trend of enrichment in the neighborhood of the start and end sites is observed for Data II. But the difference between up- and down-regulated genes is not as significant as in Data I, and their difference can only be observed in the region: -60kb upstream to 40kb intragenic and 0-60kb downstream. One possible reason for this would be the different number of binding clusters in each dataset. Because Data II contained much more binding sites than Data I, the probability is higher for the binding sites in Data II to occur in the vicinity of gene

transcription start and termination sites, even if the genes are not their targets.

To sum up for these two plots, a total of 471 genes were identified by Data II (in the sense that the region of -100kb upstream to 100kb downstream of these genes contains at least one binding sites), while 281 genes was identified by Data I. Interestingly, a high proportion of 187 genes (66.5% of 281 and 39.7% of 471 respectively) were identified by both of these two binding data. This shows a good conservation between these two binding data and raises particular interests for further analysis of these 187 genes.

As a conclusion, the binding sites are highly likely to be mapped to the neighborhood of both transcription start and termination sites. We can include these factors to construct the scoring function.

Besides calculating the number of binding sites in the vicinity of regulated genes, another way to see their association is to count the number of times that the same gene was identified by different binding sites.

3.2.2 Mapping to Binding Clusters

We sort the counts of binding clusters by genes (Data II) in this part. And given that most of the genes has only 1 to 2 binding clusters in their proximal region, there are 22 genes associated with more than 10 binding clusters (refer to Table 3.1). (totally 1786 genes,268 shows binding in the upstream 100kb distance region)

The extremely high frequencies of binding sites adjacent to genes in the Table 3.1 shows that these particular genes are strongly associated with transcription factor ER. Actually these regions are of particular biological interest (for example, NM_017679 has an alias BCAS3, which stands for breast carcinoma amplified sequence 3), and they are in the amplified region. This offers a good explanation for high number of binding sites around these genes.

Probe	Chromo	Strand	Start	End	Class	Counts
AB044555	20	+	48781730	48800432	U	16
AK093740	1	-	114239208	114248973	U	19
NM_006594	1	-	114239200	114249215	U	20
NM_015906	1	-	114741765	114855304	D	22
AF233453	20	-	45271566	45324479	D	22
NM_006526	20	-	51617018	51633043	D	22
NM_020190	1	+	114323552	114326398	U	23
AK092766	1	+	114323585	114326394	U	23
NM_014906	17	+	54188230	54417314	U	27
NM_017679	17	+	56110014	56824973	D	30
AK025510	17	+	56110037	56824980	D	30
AF010227	20	+	45645346	45715724	D	31
NM_006380	17	-	55875301	55958362	D	32
NM_183047	20	-	45271787	45418881	D	38
BX641005	20	-	45272480	45418974	D	38
AB032951	20	-	45272480	45417808	D	38
AF454056	20	-	45272511	45418850	D	38
AK000275	20	-	45272754	45418857	D	38
BC092432	20	-	45360295	45418879	D	38
BC092516	20	+	45564052	45715866	D	44
NM_006534	20	+	45564063	45719019	D	44
AF036892	20	+	45564091	45717893	D	44

Table 3.1: List of Genes which associated with more than 10 binding sites

3.2.3 Binding Associated with MoPET

***Proposition**: Binding Clusters with a larger MoPET value are more prone to be associated with regulated genes.*

According to our previous study, the number of binding clusters which are in the proximal region of regulated genes is 484 (we take the upstream region for analysis). In order to verify our hypothesis, we compare the distribution of MoPET value in the 484 binding clusters to the counterpart in the whole 4870 Clusters (with a cutoff of 11 for MoPET). Table 3.2 shows the distribution of MoPET values between Reg-Gene Associated BS and All 4870 BS.

In this table, V is the number of binding clusters which are associated with regulated genes. And E is a proportional vector of MoPET value in the whole 4870 binding clusters.

A χ^2 test can be applied to test the difference between two vectors, i.e.

$$\sum_{i=1}^{n} (E - V)^2 / E \sim \chi^2(n-1).$$

Thus the test statistic has a value of 75.7, corresponding to a p-value of 4.23×10^{-10}, which is quite significant. This show there is a shift between the two distributions vectors with an obvious accruement of percentage in the high MoPET binding clusters.

Moreover, from the Figure 4.3 we can see: when the MoPET value is less than 16, the estimated values are relatively higher; while for MoPET value over 21, the real values of associated binding clusters are comparatively larger; in between, both of the values are almost equal. Therefore, Binding Clusters with high MoPET value are more likely to be associated with regulated genes. This is in accordance with the experimental hypothesis.

MoPET No.	11	12	13	14	15	16	17	18
484 B.C (V)	85	47	29	25	19	26	19	21
Estimated Vector (E)	107	65	47	36	26	26	19	16
4870 B.C.	1076	655	477	364	258	266	187	159
MoPET No.	19	20	21	22	23	24	25	> 25
484 B.C (V)	15	10	20	12	15	13	16	112
Estimated Vector (E)	13	12	10	11	8	6	8	74
4870 B.C.	128	117	102	111	81	60	81	748

Table 3.2: MoPET Distribution in Reg-Gene Associated BS and All 4870 BS.

3.2.4 Binding Associated with Concentration

Concentration of Binding Clusters

We use windows of various length to identify those regions with high densities of binding clusters. To compare with our previous study, we map all the binding sites in the identified region to start sites of regulated genes. The results show that binding sites in the dense region obtain a relatively higher percentage in the vicinity of regulated genes.

As shown in Table 3.3, per1 measures the percentage of the number of associated binding clusters in the "windows" to the total number of binding clusters associated with regulated genes, and per2 (= 9.94%) is simply the percentage of the number of binding clusters contained in the windows out of the total 4870 binding clusters in our analysis. Per1 is slightly higher than per2 in long "windows", but the difference is more significant when the window length decreases . This suggests the concentration of binding sites may be useful for us to identify real regulation between binding sites and genes. And we can include this part to compose our scoring function.

Window Length	No. Windows	No. BS	No. BS 100kb to TSS	Per1
1kb	125	257	37	14.4%
2kb	481	1134	165	14.6%
3kb	635	1699	216	12.7%
4kb	655	2061	248	12.0%
5kb	656	2307	267	11.6%
10kb	507	2850	308	10.8%
15kb	407	3009	323	10.7%
20kb	367	3111	327	10.5%
25kb	352	3196	330	10.3%
30kb	345	3267	337	10.3%
35kb	343	3316	340	10.3%
40kb	342	3360	345	10.3%
45kb	348	3399	346	10.2%
50kb	354	3437	355	10.3%
100kb	363	3650	372	10.2%

Table 3.3: Concentration of Binding Sites.

3.3 Prediction of regulated genes using a score function

3.3.1 Score Function

Presence of a ChIP-PET binding cluster in the proximal region of a gene is not yet an evidence of transcription regulation because transcription factor binding may be related to other cellular functions or the gene to which it binds may not be really expressed [Sharov et al., 2008]. To evaluate the potential possibility of a regulated gene, we develop a score function for genes. As discussed above, we include the data of binding clusters, the distance of binding cluster to gene transcriptional starting and termination sites, MoPET and concentration of binding clusters to construct the score function. The score function is estimated as follows:

$$Score(g_i) = [\sum_{b'_j s \text{ 10kb neighborhood}} MoPET]^a * [max(min(D_{5'}, D_{3'}), 1000)/10000]^{-b}$$

where $D_{5'}$ and $D_{3'}$ are the distances of the binding cluster to $5^{'}$ and $3^{'}$ respectively.

In this score function, we make the distances to binding sites have a negative impact on the score and the summation of MoPET values have a positive impact on the score. The higher the score is, the more likely this gene is regulated. In this case, a gene will have a high score if it has very short distance to bindings sites and the MoPET values of the binding sites in its neighborhood region is high. These are in concordance with our previous findings.

We are only interested in the region of 100kb upstream to 100kb downstream, the score is set to 0 if binding cluster is out of the region. b_j is the nearest binding cluster to g_i (with the smallest $min(D_{5'}, D_{3'})$), MoPET is the maximum overlap ChIP-PET ditags and a and b are adjustable parameters. The score function is optimized to best separate between the training set of genes that were differentially

expressed in the microarray and control set of genes that were randomly selected. We use an expressed gene dataset that contains 659 up-regulated genes and 624 down-regulated genes. Adjustable parameters are changed to maximize the area of ROC (Receiver Operating characteristic) for control and expressed gene groups and the ROC curves are compared between up and down regulated genes.

3.3.2 Receiver Operating Characteristic (ROC) Curve

ROC Basics

We use the ROC curve to analyze the goodness of fit of the score function to separate genes between the control group and expressed group. After every gene is scored by our score function, we choose a cutoff to discriminate between the two groups. For those genes with score higher than the cutoff, they are classified as positive (regulated), and negative (non-regulated) otherwise. There are four cases in constructing the ROC curve (TP, FP, FN, TN):

Genes					
Test	Expressed	n	Control	n	Total
Positive	True Positive (TP)	a	False Positive (FP)	c	a + c
Negative	False Negative (FN)	b	True Negative (TN)	d	b + d
Total	a + b		c + d		

then sensitivity and specificity are defined as

$$\textbf{sensitivity} := \frac{a}{a+b}; \qquad \textbf{specificity} := \frac{d}{c+d};$$

In a ROC curve the true positive rate (Sensitivity) is plotted vs. the false positive rate (1 - Specificity) for different cut-offs [Deonier et al., 2005].

We compare the area under the ROC curve for various choices of parameters. A precise meaning of the area under an ROC curve in terms of the result of a signal detection experiment employing the two-alternative forced choice has been known for some time. [Green and Swets, 1966] showed that the area under the curve and the probability of correct classification are equal, if we assume for the moment that we have an infinite sample of observations (refers to genes in our question) that we could use the entire x continuum rather than only a finite number of category ratings. Suppose x_r and x_n stands for the score of a regulated and non-regulated gene respectively, the above conclusion can be stated as

$$\texttt{"True" area under ROC curve} = \theta = Prob(x_r > x_n)$$

And more importantly, it makes no assumptions about the form of the x_r and x_n's distributions.

ROC Curves Analysis

There are three groups of factors that can affect the plot of the ROC curve:

- parameter a and b
- different groups of binding sites : all MoPET(21047); stringent MoPET ($\geq$ 11, 4870); very stringent MoPET ($\geq$ 20, 1300)
- different choices of MoPET for score funtion :
    1. Single MoPET : only take the MoPET of the nearest binding sites to the gene of interested
    2. SiteMoPET : take all summation of MoPET for all binding sites in a particular neighborhood of the nearest binding sites
    3. GeneMoPET : take the summation of MoPET for all binding sites associated with interested gene

Analysis of ROC curve for high MoPET

After trying different combination of parameters, we could locate that the optimal choices of a and b (Table 3.4) are within the region R : $\{(a, b) : 0.5 \leq a \leq 1.5, 0.5 \leq b \leq 1.5\}$. Since a has a positive effect on the score and b has a negative effect on the score, too high of a "a" value or too low of a "b" value will highly increase the score and consequently will lead to a high false positive rate (FPR). Similarly, too low of a "a" value or too high of a "b" value will decrease the score and will lead to a high false negative rate (FNR). Both of these cases will sacrifice the accuracy of classification and reduce the area under the ROC curve.

Table 3.4 lists the values of area under ROC curve for association of high MoPET binding sites with all expressed genes versus the control genes in the region R.

As shown in the table, the area under the curve does not vary too much in this region, mostly give us a high value around $0.65 \sim 0.66$. More interestingly, if we separate the expressed genes group into up-regulated and down-regulated genes and calculate their ROC curve area respectively (listed in table 3.5 and table 3.6), up-regulated genes ($0.70 \sim 0.71$) behave much better in sense of correct identification than down-regulated genes ($0.54 \sim 0.55$), which basically is not informative.

Figure 3.4 (at a = 0.9, b = 0.7) clearly shows the difference between the ROC curve for up-regulated genes, down-regulated genes and all the genes together. This suggests that the up-regulated genes are more directly associated with binding sites, either they are much nearer to binding sites or the binding sites they associated with are of high strength.

Table 3.7 lists the area of ROC between Regulated Genes versus more Control Gene groups. From the mean value of the area, up-regulated genes are quite higher than down-regulated genes. This difference implies that ERα doesn't directly regulate down-regulated genes.

Analysis of ROC curve for SingleMoPET, SiteMoPET and GeneMoPET

We expect to see different patterns of ROC curves in the various choices of SingleMoPET, SiteMoPET and GeneMoPET. Figure 3.5 shows that the ROC curves based on SingleMoPET and SiteMoPET are quite alike (both in the shape and area). While the ROC curve based on GeneMoPET gives a lower area compared to the other two. According to our analysis in §3.2.2, some of the genes contain more than 10 binding sites in the 100kb distance. This would cause the GeneMoPET values for these genes to be extremely high and reduce the classification accuracy. In other words, due to the amplification of some of the particular regions in the ChIP-PET experiment, it is biased to take all the binding sites in 100kb to the gene to evaluate the regulation, more specifically, it may increase the FPR.

To verify, we remove all those association of binding sites with regulated genes in the amplified regions (chr1, chr3, chr8, chr17, chr20) and Figure 3.6 shows the ROC curve among various choices of MoPET after removing the amplified regions. Now all the plots clearly show that the difference between up-regulated and down-regulated genes.

3.4 Summary

We associated the gene expression data and binding data in the analysis and found that the binding strength can also help to identify the existence of regulation.

Specifically, we have shown that binding clusters with higher MoPET values are more likely to be associated with regulated genes and the binding clusters enriched-region also showed a stronger association with regulated genes. To integrate of all these findings, we defined a score function for genes which included these important factors. Under these metric, potential regulated genes should score higher than non-regulated genes. The score function can help us identify regulated genes in separating expressed and control gene groups. Also, it may help to assess different

groups of expressed genes. Accuracy of the score function to separate expressed and control genes was evaluated by ROC curve analysis. A number of parameters choices have been tested for the ROC curve and the numerical results showed the preference of regulation to those genes which are associated with high MoPET, but only for up-regulated genes.

a/b	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0.5	0.663	0.660	0.656	0.654	0.652	0.650	0.649	0.645	0.647	0.645	0.643
0.6	0.665	0.663	0.659	0.658	0.654	0.652	0.651	0.653	0.649	0.647	0.642
0.7	0.667	0.664	0.663	0.661	0.659	0.656	0.654	0.652	0.650	0.648	0.646
0.8	0.668	0.667	0.664	0.663	0.661	0.659	0.656	0.654	0.651	0.650	0.649
0.9	0.668	0.668	0.666	0.664	0.663	0.661	0.657	0.656	0.654	0.652	0.650
1	0.668	0.668	0.667	0.665	0.661	0.660	0.659	0.657	0.654	0.652	0.651
1.1	0.670	0.667	0.667	0.665	0.662	0.659	0.659	0.658	0.656	0.653	0.649
1.2	0.666	0.666	0.663	0.663	0.659	0.658	0.656	0.654	0.653	0.652	0.651
1.3	0.656	0.659	0.659	0.660	0.658	0.657	0.655	0.654	0.654	0.654	0.652
1.4	0.649	0.649	0.647	0.651	0.649	0.647	0.652	0.650	0.649	0.649	0.648
1.5	0.619	0.630	0.633	0.631	0.631	0.633	0.637	0.636	0.634	0.636	0.636

Table 3.4: Table of Area under ROC curve of All Regulated Genes vs. Control Genes by Parameter a&b(high MoPET $\geq$20)

a/b	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0.5	0.707	0.705	0.702	0.700	0.697	0.697	0.695	0.692	0.693	0.691	0.689
0.6	0.708	0.706	0.704	0.704	0.701	0.698	0.698	0.698	0.694	0.692	0.687
0.7	0.710	0.708	0.707	0.706	0.704	0.702	0.700	0.697	0.696	0.693	0.691
0.8	0.710	0.710	0.708	0.706	0.706	0.704	0.702	0.699	0.697	0.696	0.696
0.9	0.709	0.711	0.710	0.708	0.706	0.705	0.701	0.701	0.700	0.698	0.696
1	0.710	0.710	0.709	0.709	0.704	0.703	0.703	0.701	0.700	0.698	0.697
1.1	0.711	0.709	0.708	0.707	0.705	0.702	0.702	0.702	0.700	0.698	0.694
1.2	0.706	0.706	0.703	0.705	0.701	0.700	0.699	0.696	0.696	0.695	0.695
1.3	0.692	0.697	0.700	0.701	0.699	0.699	0.698	0.696	0.696	0.697	0.696
1.4	0.683	0.685	0.685	0.690	0.687	0.688	0.693	0.692	0.691	0.691	0.691
1.5	0.647	0.662	0.668	0.668	0.668	0.671	0.678	0.677	0.675	0.678	0.678

Table 3.5: Table of Area under ROC curve of Up-regulated Genes vs. Control Genes by Parameter a&b(high MoPET $\geq$20)

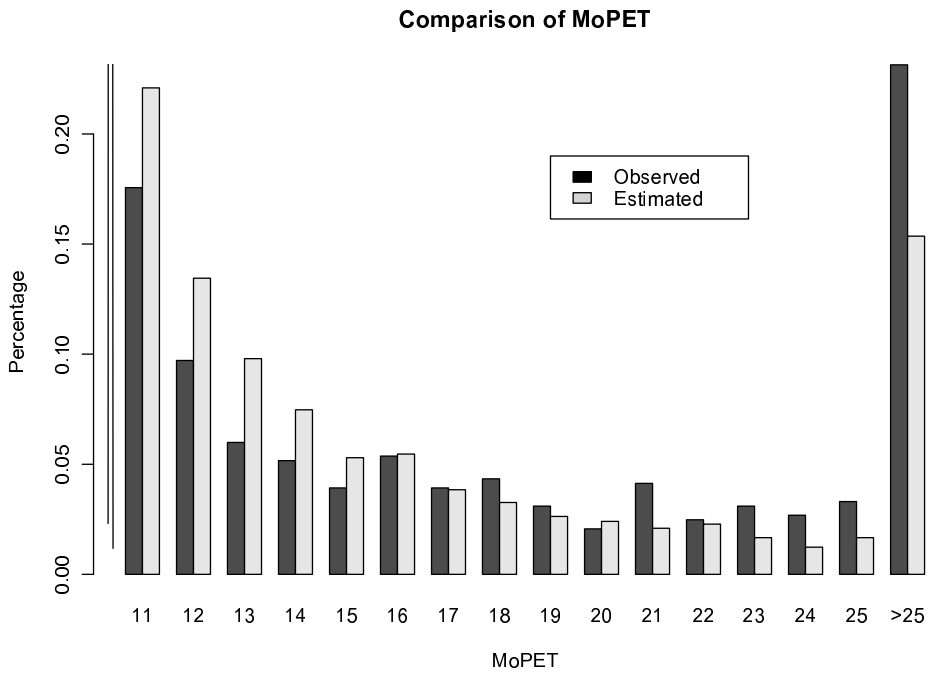

Figure 3.3: Comparison of MoPET in Reg-Gene Associated BS and All 4870 BS

a/b	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
0.5	0.570	0.564	0.560	0.557	0.556	0.553	0.554	0.548	0.550	0.548	0.547
0.6	0.573	0.572	0.564	0.562	0.557	0.555	0.553	0.558	0.554	0.553	0.546
0.7	0.577	0.570	0.571	0.568	0.563	0.558	0.558	0.556	0.553	0.552	0.550
0.8	0.581	0.577	0.571	0.571	0.567	0.564	0.559	0.559	0.554	0.552	0.552
0.9	0.580	0.579	0.575	0.570	0.570	0.567	0.563	0.561	0.556	0.557	0.552
1	0.581	0.581	0.577	0.574	0.569	0.569	0.566	0.563	0.559	0.556	0.555
1.1	0.584	0.579	0.581	0.576	0.571	0.568	0.568	0.565	0.563	0.559	0.553
1.2	0.584	0.580	0.578	0.577	0.571	0.569	0.565	0.564	0.562	0.560	0.558
1.3	0.581	0.579	0.574	0.575	0.572	0.569	0.566	0.564	0.565	0.564	0.561
1.4	0.578	0.572	0.568	0.570	0.567	0.562	0.565	0.561	0.560	0.560	0.558
1.5	0.558	0.562	0.561	0.554	0.553	0.552	0.553	0.549	0.548	0.549	0.549

Table 3.6: Table of Area under ROC of Down-regulated Genes vs. Control Genes by Parameter a&b(high MoPET $\geq$20)

	R1	R2	R3	R4	R5	R6
All	0.667	0.631	0.652	0.666	0.621	0.646
Up-reg	0.711	0.677	0.692	0.714	0.664	0.691
Down-reg	0.576	0.534	0.566	0.566	0.527	0.55

	R7	R8	R9	R10	mean	variance
All	0.673	0.655	0.687	0.625	0.6523	0.0218
Up-reg	0.721	0.702	0.729	0.667	0.6968	0.0225
Down-reg	0.573	0.557	0.6	0.535	0.5584	0.0225

Table 3.7: Area under ROC Curve for Expressed vs. 10 Control Groups (MoPET $\geq$ 20)

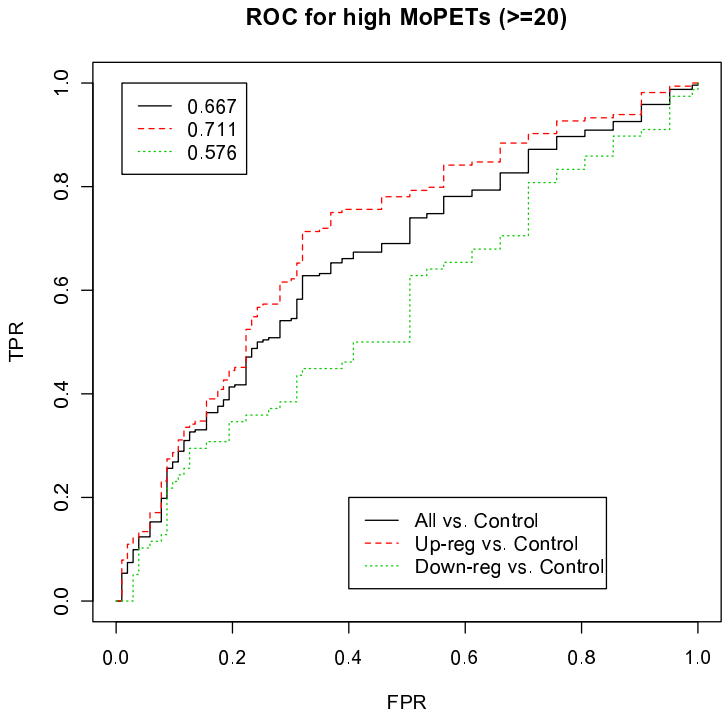

Figure 3.4: ROC curve for high MoPET($\geq$ 20) at a = 0.9, b = 0.7 with Single-MoPET

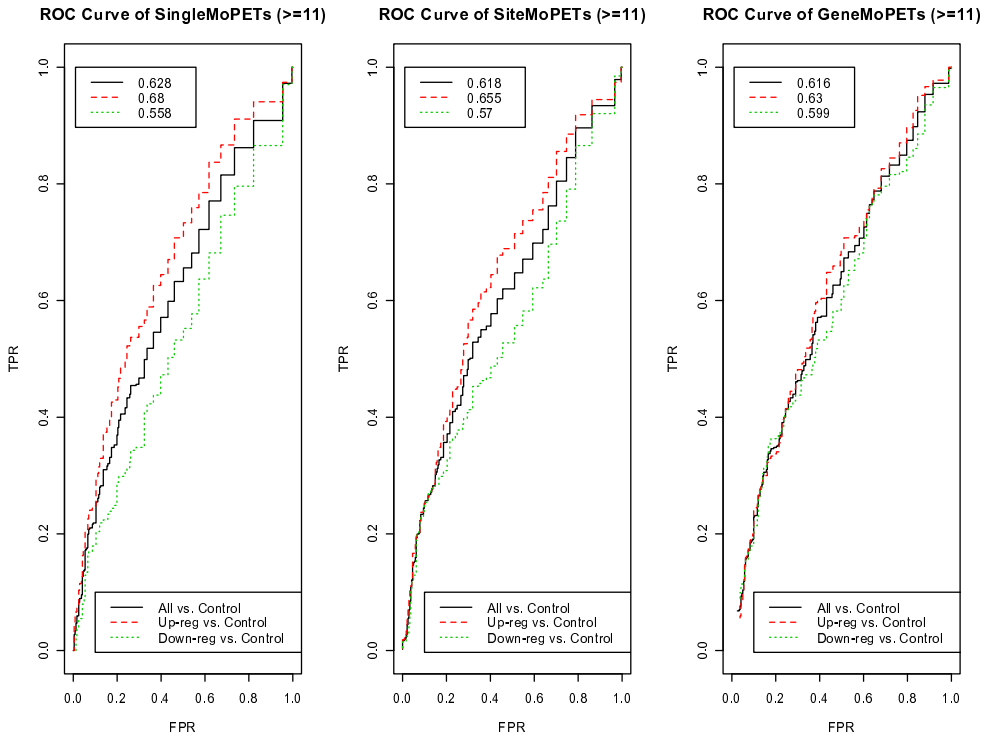

Figure 3.5: Comparison of ROC curve for Stringent MoPET($\geq$ 11) with Single-MoPET, SiteMoPET, GeneMoPET

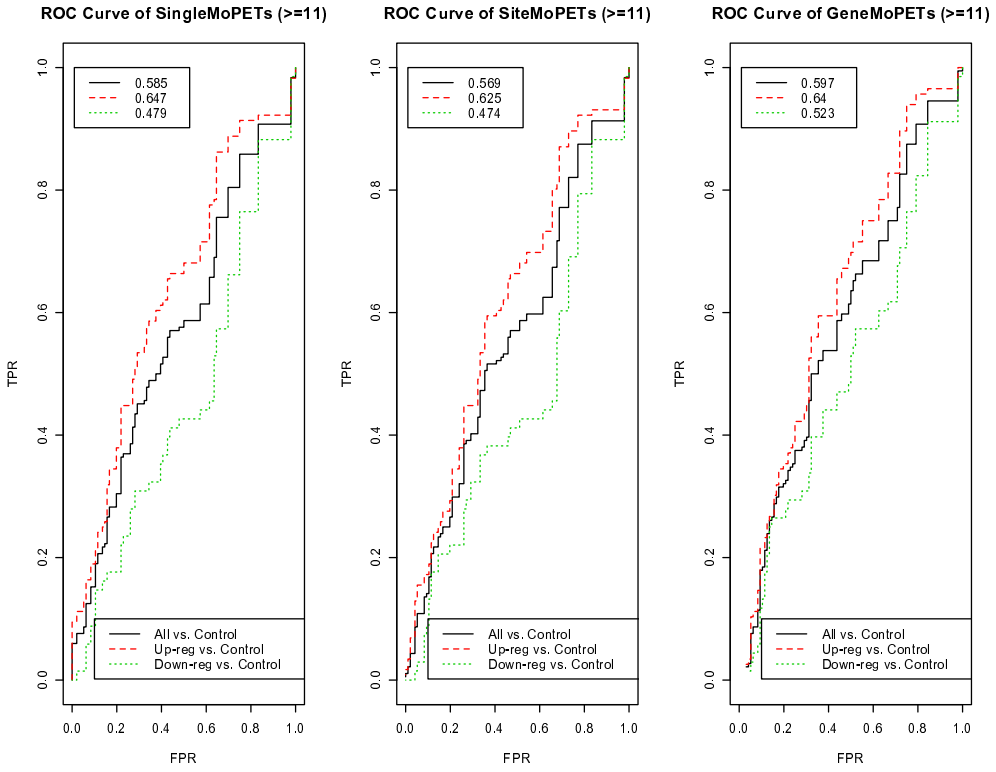

Figure 3.6: Comparison of ROC curve for Stringent MoPET($\geq$ 11) with SingleMoPET, SiteMoPET, GeneMoPET after removing amplified regions(Chr1,3,8,17,20)

Chapter 4

Discussion

The identification of targets of a transcriptional factor such as the estrogen receptor across the whole genome provides an important new source for the study of gene regulation. The classic paradigm of estrogen receptor function involves binding to promoter-proximal regions and subsequent gene regulation. However, it now seems that the promoter-proximal region, although important for some genes, do not constitute the majority of estrogen receptor target sites [Lin et al., 2007].

Our proposal was to integrate various datasets and explore the gene expression data. Our data-driven analysis allows us to test various mechanistic hypotheses about what the rules for gene regulation might be. We have already tested the distance, binding strength and concentration of binding regions, and have shown that these factors were important in different degrees. Tentatively we proposed a score function for genes to measure their potential to be directly regulated by including these factors. The numerical results between control gene group and expressed gene group were shown and compared by the Receiver Operating Characteristic (ROC) curve analysis.

However, because the exact differentially expressed genes are unknown, we cannot

verify our results in the biological sense. And with limited information, the score function can only be used to differentiate two groups of genes, not individual genes.

In the thesis we have only considered to divide regulated gene groups into up and down regulated groups. Generally, we observed that generally the up-regulated genes scored higher than down-regulated genes. Rather than simply divide the genes into a binary up and down classification, in future we could explore ways for the grouping to identify more refined groups of genes that behave in a consistent way after the ER binding.

Another aspect for future work is that we can extend our work to other kinds of TFs. Different TFs will have different mechanisms of gene regulation. For instance, apart from activator proteins such as ERα, we might look at insulator proteins (e.g. CTCF) which are thought to create regulatory boundaries [Bell et al., 1999]. In addition, it would be interesting to study models combining multiple TF datasets. For example, two ES proteins, Oct4 and Sox2, act together in Embryonic Stem Cells [Chen et al., 2008].

Bibliography

[Bell et al., 1999] Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein ctcf is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98:387–396.

[Buck and Lieb, 2004] Buck, M. J. and Lieb, J. D. (2004). Chip-chip: considerations for the design, analysis, and the application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83:349–360.

[Carroll et al., 2006] Carroll, J. S., Meyer, C. A., and Song, J. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics*, 38(11):1289–1297.

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

[Deonier et al., 2005] Deonier, R. C., Tavare, S., and Waterman, M. S. (2005). *Computational Genome Analysis*. Springer.

[Green and Swets, 1966] Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.

[Irizarry et al., 2003] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15.

[Lee, 2004] Lee, M.-L. T. (2004). *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers.

[Lin et al., 2007] Lin, C.-Y., Vega, V. B., Thomsen, J. S., Zhang, T., Kong, S. L., Xie, M., Chiu, K. P., Lipovich, L., H.Barnett, D., Stossi, F., Yeo, A., George, J., Kuznetsov, V. A., Lee, Y. K., Charn, T. H., Palanisamy, N., Miller, L. D., Cheung, E., Katzenellenbogen, B. S., Ruan, Y., Bourque, G., Wei, C.-L., and Liu, E. T. (2007). Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genetics*, 3(6):867–885.

[Metz, 1978] Metz, C. E. (1978). Basic principles of roc analysis. *Seminar in Nuclear Medicine*, 8(4):283–298.

[Parmigiani et al., 2003] Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L., editors (2003). *The Analysis of Gene Expression Data*. Statistics for Biology and Health. Springer.

[Sharov et al., 2008] Sharov, A. A., Masui, S., Sharova, L. V., Piao, Y., Aiba, K., Matoba, R., Xin, L., Niwa, H., and Ko, M. S. (2008). Identification of pou5f1, sox2, and nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*, 9:269.

[Storey, 2002] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498.

[Tusher et al., 2001] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science of the United States of America*, 98(9):5116–5121.

[Wei et al., 2006] Wei, C.-L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X., Chew, J.-L., Lee, Y. L., Kuznetsov, V. A., Sung, W.-K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H.-H., and Ruan, Y. (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124:207–219.