Bayesian Modeling for High Throughput Genomic Data

by

Ming Hu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Assistant Professor Zhaohui Qin, Chair
Professor Gonçalo Abecasis
Professor Jeremy M. G. Taylor
Assistant Professor Jiandie Lin
Research Associate Professor Timothy D. Johnson
Research Assistant Professor Chandan Kumar

ACKOWLEDGEMENTS

I am most grateful to my advisor, Dr. Zhaohui Steve Qin, for his invaluable supervision of and guidance on this thesis. I am also grateful to Drs. Jeremy M. G. Taylor, Gonçalo Abecasis, Jiandie Lin, Timothy D. Johnson and Chandan Kumar for their helpful advice and useful comments.

The work presented in this thesis would not have been possible if it were not for Dr. Yu Michael Zhu, who introduced me into spatial statistics, and Dr. Jun Li, who provided me RNA-Seq data. Dr. Arul Chinnaiyan's lab generously allowed me to use their data in the development of the motif finding program. Special thanks go to the referees at PLoS One and Nucleic Acid Research for their comments that led to valuable improvement of the research.

This thesis is dedicated to my parents. I would like to thank them for their support and love. It is only because of them that I understand the values of my professional career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Background

The explosion of high throughput genomic data in recent years has already altered our view of the extent and complexity of biology. In the past decade, microarray technology (1) has played a prominent role in advancing our understanding of transcriptome complexity. Microarray is a hybridization-based technology that incubates fluorescently-labeled cDNA with custom-made microarray or commercial GeneChips. Microarray allowed scientists to simultaneously monitor the expression of almost all the genes in the genome, and along with a steady reduction in processing costs, led to its wide spread application. ChIP-chip (2), a technology that combines chromatin immunoprecipitation with microarray, has been widely utilized to investigate interactions between proteins and DNA *in vivo*. In ChIP-chip experiments, the protein-DNA binding is recognized by detecting hybridization signals using a fixed set of probes on DNA microarrays. However, due to their restriction of the probes present on the DNA microarray, such methods are naturally limited in scale and resolution. Whole-genome tiling arrays are also cost prohibitive and technically challenging.

With the recent development of next generation sequencing, array-based technologies have being replaced by sequencing-based technologies, such as RNA-Seq (3) and ChIP-Seq (4), which can dramatically improve the quantity and quality of high throughput genomic data. In RNA-Seq experiments, a population of RNA is converted to a library of cDNA fragments with adaptors attached to one end. Each molecule, after amplification, is then sequenced using next generation sequencing technologies. Following sequencing, the resulting reads are aligned to either the reference genome or known transcripts to produce a genome-scale transcriptional profile. Meanwhile, ChIP-Seq, or chromatin immunoprecipitation followed by ultra-high-throughput sequencing, has emerged as a powerful new technology for genome-wide mapping of protein-DNA interactions and histone modifications. Through direct sequencing of all DNA fragment from ChIP assays, ChIP-Seq is capable of revealing protein-DNA interaction sites across the entire genome, thus building a comprehensive and high-resolution interactome map for DNA-binding proteins of interest. The comprehensiveness and the high resolution are two key advantages of ChIP-Seq over ChIP-chip.

To analyze such high throughput genomic data is complicated. While sequencing cost has significantly declined, the amount of data the new platforms produce is skyrocketing, thereby producing an analytical bottleneck. Detecting biological signals from experimental noises is similar to looking for a needle in a haystack. Technologically specific features, heterogeneous data structures and massive sample sizes present great challenges and opportunities to develop novel statistical methodologies in computational biology.

Several analytical modalities have been proposed for analyzing high throughput genomic data, such as Boolean networks (5), gene networks (6) and genetic algorithms (7). A majority of them are frequentist modeling methods which involve complicated inference procedure and unstable numerical implementation. In this context, a Bayesian modeling method appears to be an attractive alternative.

In this dissertation, we describe three Bayesian modeling methods in high throughput genomic data analysis. The general Bayesian modeling procedure can be described as consisting of two main steps (8): (a) setting up a full probability model, the joint distribution $f(y|\theta)$, that captures the relationship among all the variables (e.g., observed data y, unknown parameters θ) into consideration; (b) summarizing the findings for particular quantities of interest using appropriate posterior distributions, which is typically a conditional distribution of the quantities of interest given the observed data.

A standard procedure for carrying out step (a) is to formulate the scientific question of interest though the use of a probabilistic model from which we can write down the likelihood function of unknown parameters. In the analysis of high throughput genomic data, this step involves understanding the underlying biological process and the data generation mechanism. Then a prior distribution $f_0(\theta)$ is contemplated, which should be both mathematically tractable and scientifically meaningful. Usually $f_0(\theta)$ is derived from previous biological findings. The joint probability distribution can then be represented as the product of the likelihood and the prior $p(y, \theta) = p(y|\theta)f_0(\theta)$. Step (b) is completed by obtaining the posterior distribution through the application of Bayes theorem:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)f_0(\theta)}{\int p(y|\theta)f_0(\theta)d\theta} \propto p(y|\theta)f_0(\theta)$$

After integrating out nuisance parameters, we get full conditional distribution of unknown parameter θ and conduct the standard posterior sampling based inference via Markov Chain Monte Carlo techniques.

The Bayesian modeling approach has at least two advantages in analyzing high throughput genomic data. First, through the prior distribution, we can use prior knowledge and information about the value of unknown parameters. This is especially important since biologists often have substantial knowledge about the subject under study. To the extent that this information is correct, it will sharpen the inference about the unknown parameters and accelerate the detection of biological signals. Second, treating all the variables in the system as random variables greatly clarifies the methods of analysis. It follows from the basic probability theory (Bayes formula) that information about the realized value of any random variable based on observation of related random variables is summarized in the conditional distribution.

1.2 Outline

This thesis consists of three chapters that each addresses an independent statistical problem in high throughput genomic data analysis.

In chapter 2, we develop a querying algorithm for analyzing microarray compendium data (9). In microarray gene expression data analysis, it is often of interest to identify genes that share similar expression profiles with a particular gene such as a key regulatory protein. While working well for small datasets, the heterogeneity introduced from increased sample size inevitably reduces the sensitivity and specificity of these

approaches. This is because most co-expression relationships do not extend to all experimental conditions. With the rapid increase in the size of microarray datasets, identifying functionally related genes from large and diverse microarray gene expression datasets is a key challenge. We develop a model-based gene expression query algorithm built under the Bayesian model selection framework that is capable of detecting co-expression profiles under a *subset* of samples/experimental conditions. In addition, this algorithm allows linearly transformed expression patterns to be recognized and is robust against sporadic outliers in the data. Both features are critically important for increasing the power of identifying co-expressed genes in large scale gene expression datasets. Our simulation studies suggest that this method outperforms existing correlation coefficients or mutual information-based query tools. When we apply this new method to the *Escherichia coli* microarray compendium data, it identifies a majority of known regulons, as well as novel potential target genes of numerous key transcription factors.

In chapter 3, we study the *de novo* motif finding problem using ChIP-Seq data (10). Coupling chromatin immunoprecipitation (ChIP) with recently developed massively parallel sequencing technologies has enabled genome-wide detection of protein–DNA interactions with unprecedented sensitivity and specificity. This new technology, ChIP-Seq, presents opportunities for in-depth analysis of transcription regulation. In this study, we explore the value of using ChIP-Seq data to better detect and refine transcription factor binding sites (TFBS). We introduce a novel computational algorithm named Hybrid Motif Sampler (HMS), specifically designed for TFBS motif discovery in ChIP-Seq data. We propose a Bayesian model that incorporates sequencing depth information to aid in motif identification. Our model also allows intra-motif dependency to describe

more accurately the underlying motif pattern. Our algorithm combines stochastic sampling and deterministic comprehensive search steps into a novel hybrid iterative scheme. This combination accelerates the computation process. Simulation studies demonstrate favorable performance of HMS compared to other existing methods. When applying HMS to real ChIP-Seq datasets, we find that (i) the accuracy of existing TFBS motif patterns can be significantly improved; and (ii) there is significant intra-motif dependency inside all the TFBS motifs we tested; modeling these dependencies further improves the accuracy of these TFBS motif patterns. These findings may offer new biological insights into the mechanisms of transcription factor regulation.

In chapter 4, we design a model-based gene expression measurement using RNA-Seq data. High throughput sequencing technology, also called RNA-Seq, has become a revolutionary tool for transcriptomics analysis. Compared to microarray, RNA-Seq offers clear advantages, such as better dynamic range and the ability to discover novel transcripts. Current methods enumerate the number of reads within each exon and use the normalized read counts to represent the gene expression levels for that exon. However, the sequencing depth across exons fluctuates substantially and shows significant spatial correlation. These variations will affect the simple enumeration method that is currently being used. In this study, we propose a spatial Poisson regression model to provide a portrait of base-level sequencing depth within each exon. This method can capture local genomic features that affect coverage depth in the spatial model, and therefore, offer improved quantification of the true underlying expression levels.

CHAPTER 2

Query large scale microarray compendium datasets using a model-based Bayesian approach with variable selection

2.1 Introduction

Genome-wide expression analysis with DNA microarray technology (1),(11). has become an indispensable tool in genomics research (12). Increased accessibility, lowered cost and improved technology result in more comprehensive studies, under more diverse conditions and a rapid expansion of available gene expression data. This presents an important resource for mining biological information. A particular example is the so-called microarray compendium in which gene expression profiles were surveyed in hundreds of samples which were treated under diverse biological conditions (13-15). Data generated from such studies is highly informative. However, due to heterogeneity, finding biological insight from such datasets proves a major challenge. Scalable and effective mining tools capable of extracting knowledge from diverse and noisy information sources are critically needed (16).

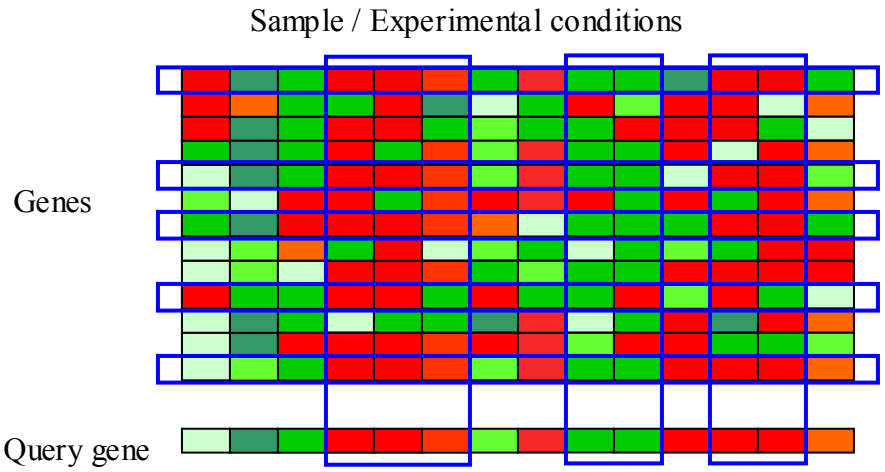
An effective data mining tool for gene expression microarray data is to infer relatedness among genes based on their expression profile, a tactic referred to as the "guilt by association" (GBA) principle (17-21). The underlying hypothesis is that

functionally related genes, such as transcription factor (TF) and its regulated genes—regulon —tend to display correlated gene expression patterns. For example, Mootha et al. (2003) (22) proposed the "neighborhood analysis algorithm" to identify "neighboring" genes that share correlated expression profiles with genes of interest. Various measurements such as Pearson correlation, Spearman's rank correlation, Kendall's τ and mutual information (23) have been used to assess the strength of the correlation. Recently, much interest has been generated on genome-wide regulatory network inference (24), where pairwise regulatory relationships among genes need to be predicted. As an example, Faith et al. (2007) (13) developed the context likelihood of relatedness (CLR) algorithm to identify regulatory interactions.

Although successful in analyzing small datasets, the above mentioned correlation or distance measures will be less helpful for searching large datasets, such as microarray compendium data. This is because for most functionally related genes, tight correlation only occurs under specific experimental conditions. Therefore global correlation measures taken across diverse experimental conditions will be significantly reduced, and thus make it harder to recognize functional related genes. Given the microarray compendium scenario, we hypothesized that statistically significant correlation can still be detected using microarray, but strong correlation will be confined to a subset of samples/experimental conditions. Under this hypothesis, it is highly desirable to develop a query tool that can automatically recognize a subset of conditions under which the query gene and its targets share tightly correlated expression profiles. This is analogous to the development of local alignment tools such as BLAST (25) to search for subtle patterns in large amounts of sequence data.

In this study, we design a model-based query algorithm capable of detecting significantly correlated expression patterns that are restricted to a subset of experimental conditions. See Figure 2.1 for an illustration of our scheme. This approach not only predicts functionally related genes, it also allows one to discover under which experimental conditions such co-expression occurs. The proposed query tool will provide researchers with a much needed device to explore the rich resources of vast microarray databases available. This model is inspired by the Bayesian Partition with Pattern Selection (BPPS) model designed to identify functionally related proteins (26). Our proposed method is related to bi-clustering (27-31) since we consider both genes and samples/experimental conditions. However, bi-clustering is unsupervised, which is different from the supervised pattern matching procedure we propose. Qian et al. (2001) (32) introduced a pairwise query algorithm for gene expression data based on a Smith-Waterman type local alignment algorithm (33). However, that algorithm is designed for querying time-course gene expression data only, and is generally not applicable to datasets where the experimental conditions are unrelated. Dhollander et al. (2007) (34) introduced a model-based query-driven module discovery tool—QDB, but it is aimed at performing informed bi-clustering instead of pattern matching, and it does not take into account the complex correlation patterns such as inverse patterns. Owen et al. (2003) (35) proposed a score-based search algorithm called gene recommender (GR) to find genes that are co-expressed with a given set of genes using data from large microarray datasets. GR first selects a subset of experiments in which the query genes are most strongly co-regulated. Hence multiple query genes are required.

Figure 2.1 Illustration of the model-based gene expression query algorithm. Each row represents a gene, and each column represents a sample/experimental condition. The query gene is at the bottom. The Blue boxes indicate the collection of genes and experimental conditions in which co-expression with the query gene is observed.



2.2 Methods

2.2.1 Statistics model

We propose a model-based query tool for gene expression data. The goal is to identify genes that share correlated expression profiles with a particular gene such as a key TF.

The entire microarray compendium can be represented as a matrix, where each row represents a gene and each column represents an experimental condition. We are hoping to identify a subset of rows (genes) and a subset of columns (conditions) such that these genes show co-expression with the query gene under the selected conditions. This procedure is similar to placing binary labels on all rows and columns. Finding the maximum likelihood estimator is often a good solution to such a statistical inference problem. However, the large number of rows and columns make it impossible for us to

enumerate all possible combinations. We therefore employ a Markov Chain Monte Carlo strategy to guide an efficient search. The statistical model and computational algorithm is as follows (more details can be found in the appendix).

Suppose there is a database containing expression levels of N genes across M different experimental conditions. Each gene is represented by an expression vector $y_i = (y_{i1}, y_{i2}, \ldots, y_{iM})$ that can be summarized as a data matrix $Y = (y_1, y_2, \ldots, y_N)^t$, Given a particular query expression profile $x = (x_1, x_2, \ldots, x_M)$, we want to identify all genes that share similar expression patterns across a subset of experimental conditions. To do this, we define a difference vector as $z_i = (y_{i1} - x_1, y_{i2} - x_2, \ldots, y_{iM} - x_M)$, and use $z = (z_1, z_2, \ldots, z_N)^t$ as the input data for our inference. We introduce a row indicator vector $R = (r_1, r_2, \ldots, r_N)$ and a column indicator vector $E = (e_1, e_2, \ldots, e_M)$, $r_i = 1$ indicates that gene i in the database is functionally related to the query gene and 0 otherwise. $e_j = 1$ indicates that co-expression occurs at the j th experimental condition (foreground) and 0 otherwise (background). We assume that the differences between a related gene and the query gene at the foreground columns follow normal distributions $z_{ij} \sim N(0, \sigma_{1j}^2)$. The remainder of Z is assumed to follow background normal distributions $z_{ij} \sim N(\mu_{0j}, \sigma_{0j}^2)$ where $\sigma_{1j}^2 < \sigma_{0j}^2$. Let $\phi(x|\mu, \sigma^2)$ represents the probability density function of normal distribution with mean μ and variance σ^2. The overall likelihood can be expressed as:

$$P(Z|R, E, \Theta) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\phi(z_{ij}|0, \sigma_{1j}^2)^{r_i \cdot e_j} \phi(z_{ij}|\mu_{0j}, \sigma_{0j}^2)^{1 - r_i \cdot e_j} \right] \qquad (2.1)$$

where $\Theta = (\mu_{01}, \mu_{02} \ldots, \mu_{0M}, \sigma_{01}^2, \sigma_{02}^2, \ldots, \sigma_{0M}^2, \sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{1M}^2)$. We adopt standard conjugate priors for these model parameters (8).

$$p(\sigma_{1j}^2) \sim Inverse\ Gamma(\alpha_{1j}, \beta_{1j}),$$

$$p(\sigma_{0j}^2) \sim Inverse\ Gamma(\alpha_{0j}, \beta_{0j}),$$

$$p(\mu_{0j} | \sigma_{0j}^2) \sim N(\tau_{0j}, \sigma_{0j}^2),$$

$$p(r_i) \sim Bernoulli(\pi_{r_i}),$$

$$p(e_j) \sim Bernoulli\left(\pi_{e_j}\right),$$

$$i = 1, \ldots, N, \quad j = 1, \ldots, M. \tag{2.2}$$

Specification of these prior distributions can be found in the appendix.

Parameters of interest are the two indicator vectors R and E. Θ is regarded as the nuisance parameter and is integrated out to simplify the computation (36). We use the Gibbs sampler (37,38) to sample R and E from the posterior distributions. To be specific, our algorithm will cycle through all rows and columns sequentially, flip the indicator variables of each row or column, and then decide whether to accept the change based on the Bayes factor calculated. The joint distributions can be derived as follows:

$$P(R, E, \Theta | Z) \propto P(Z | R, E, \Theta) \prod_{i=1}^{N} p(r_i) \prod_{j=1}^{M} \left[p(\mu_{0j} | \sigma_{0j}^2) p(\sigma_{0j}^2) p(\sigma_{1j}^2) p(e_j) \right] \tag{2.3}$$

After integrating out nuisance parameters, we get full conditional distributions of R and E, which are Bernoulli distributions. The details can be found in the appendix.

The detailed procedure of our algorithm is as follows.

(1) Initialization: randomly assign row and column indicators to be either one or zero. Calculate the differences $z_{ij} = y_{ij} - x_j$.

(2) Cycle through all rows and columns sequentially 50 times. At each cycle, draw the indicator for each row and column from the full conditional distributions. The result with the highest log likelihood during the 50 cycles is recorded.

(3) Repeat the cycle ten times, and report the result with the highest log likelihood from all runs.

In the initialization step, the row and column indicators can be assigned randomly. In practice, one can simply assign 1 to the top half of rows and to the first half of columns and 0 to the rest of rows and columns. If there is additional information suggesting certain genes (rows) are targets (or non-targets) of the TF, it is recommended that the indicator 1 (or 0) be assigned to those genes and the same for the experimental conditions.

2.2.2 Add linear factor

In the previous model, we require that the target genes and the query gene share similar expression levels in selected experimental conditions. This is restrictive since functionally related genes may display the same expression pattern but differ in absolute quantity. To capture this, we extend our model to allow the expression levels of the target gene and the query gene to differ by a constant factor. That is, their expression profiles are proportional to each other: $y_{ij} = a_i x_j$. Here a_i is a linear transformation factor for gene i. a_i can be either positive or negative indicating positive or negative correlation respectively. After normalization, we estimate the linear transformation factor a_i using least square without intercept. To keep our model simple and avoid over-fitting, we restrict the linear factor to be significantly different from 0. The estimation step is made at the beginning of each cycle based on the most recently updated column indicators.

2.2.3 Allow cell-level noise

In the aforementioned models, genes selected are mandated to have similar expression profiles up to a constant factor under a subset of experimental conditions. Hence the chosen rows and columns in the original data matrix form a solid block when combined. This may still be too restrictive because a few sporadic cells in the block may deviate from the corresponding values in the query profile. Possible reasons that may cause such discrepancy are experimental artifacts, measurement errors, or substructures in the co-expression pattern. To account for this, we introduce an additional binary indicator variable, c_{ij}, for each cell in this block to indicate whether this particular gene/experimental condition combination should be treated as background. This additional step allows us to identify significant but imperfect patterns. Adding this additional parameter, the overall likelihood is modified as follows:

$$P(Z|R,E,C,\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\phi(z_{ij}|0,\sigma_{1j}^2)^{r_i\cdot e_j\cdot c_{ij}}\phi(z_{ij}|\mu_{0j},\sigma_{0j}^2)^{1-r_i\cdot e_j\cdot c_{ij}}\right] \quad (2.4)$$

We use a Bernoulli distribution as the prior for c_{ij},

$$p(c_{ij}) \sim Bernoulli\left(\pi_{c_{ij}}\right), i = 1,\dots,N, j = 1,\dots,M, \quad (2.5)$$

The prior for this new indicator variable will be set such that only a small fraction of cells is allowed to be treated as background.

After integrating out nuisance parameters, the full conditional distributions of all model parameters can be obtained similarly as before. The details can be found in appendix.

2.3 Results

The aforementioned algorithm has been implemented in a C++ program named BEST (Bayesian Expression Search Tool). To test its performance, we applied it to a series of synthetic datasets as well as to the real *Escherichia coli* microarray compendium dataset (39). In addition to BEST, we also tested well-established query tools based on Pearson, Spearman correlation coefficients, Kendall's τ, mutual information (23) and the model-based query-driven module discovery tool—QDB (34).

2.3.1 Synthetic datasets

All simulated data contained 100 rows (genes) and 50 columns (experimental conditions). Around 20% of the 100 genes were randomly assigned as the "target" genes. Let T represent the total number of target genes in a dataset. To mimic the scenarios that gene expression correlation only presents in a subset of experimental conditions, we separated the 50 columns into foreground and background and require that correlated expression profiles between the query gene and the target genes can only be observed among foreground columns. To assess the impact of the proportion of foreground columns on the effectiveness of identifying target genes, we tested four different settings: 100%, 75%, 50% and 25% of columns were selected as foreground. At each foreground column, the expression profiles of the query gene and T target genes were generated from a $T + 1$ dimensional multivariate normal distribution with mean zero and variance-covariance matrix Σ. The correlation coefficient between the query gene and each target gene was set to be 0.95.

The remaining expression profiles were generated independently from a uniform distribution between -4 and 4. To mimic the noisy nature of the microarray data, we

included the following additional settings: randomly add linear transformations to 50% of the target genes (the linear transformation factors were randomly picked from $(-2, -1, -0.5, 0.5, 2)$; randomly add additional noise (± 5) to 10% of the expression values of target genes in foreground columns to mimic outliers caused by experimental artifacts. We also considered settings in which neither or both of these two complications were present. The combination of these four scenarios with the four different proportions of foreground columns mentioned above resulted in 16 different testing cases. We generated 50 simulated datasets for each of the 16 cases, and tested all query methods on each dataset to identify target genes. To compare performance, we sorted the 100 genes using the relatedness measures adopted in each method and found the proportions of true positives among the top T genes. The means and standard deviations of these proportions were summarized in Table 2.1. We also produced Receiver Operating Characteristic (ROC) curves for all methods under all simulation settings. ROC curves obtained from the most challenging scenario, where only 25% of the columns are foreground, are shown in Figure 2.2. ROC curves obtained from other simulation settings can be found in Figures A2.1, A2.2 and A2.3 in appendix. The areas under the curve (AUC) of these ROC curves were summarized in Table A2.1 in appendix.

Table 2.1 Performance comparison among various methods for querying simulated microarray gene expression dataset. Best results are displayed in bold.

Case	Sub-case*	Pearson[a]	Spearman[b]	Kendall[c]	QDB[d]	Mutual[e]	BEST A[f]	BEST B[g]	BEST C[h]
Case 1:	I	**1 (0)**[1]	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**
100%	II	0.67 (0.12)	0.68 (0.12)	0.68 (0.12)	0.59 (0.13)	**1 (0.01)**	**1 (0)**	**1 (0)**	**1 (0)**
foreground	III	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**	**1 (0)**
	IV	0.62 (0.09)	0.70 (0.09)	0.70 (0.09)	0.51 (0.11)	0.78 (0.08)	0.97 (0.04)	**0.98 (0.03)**	**0.98 (0.03)**
Case 2:	I	0.89 (0.10)	0.96 (0.05)	0.99 (0.03)	**1 (0)**	0.87 (0.09)	**1 (0)**	**1 (0)**	**1 (0)**
75%	II	0.66 (0.12)	0.71 (0.10)	0.70 (0.09)	0.70 (0.10)	0.81 (0.09)	**1 (0)**	**1 (0)**	**1 (0)**
foreground	III	0.91 (0.09)	0.97 (0.04)	0.99 (0.03)	**1 (0)**	0.87 (0.09)	**1 (0)**	**1 (0)**	**1 (0)**
	IV	0.61 (0.11)	0.68 (0.11)	0.70 (0.11)	0.53 (0.12)	0.70 (0.11)	**0.97 (0.04)**	**0.97 (0.04)**	**0.97 (0.04)**
Case 3:	I	0.66 (0.17)	0.73 (0.14)	0.80 (0.13)	0.97 (0.16)	0.61 (0.14)	**1 (0)**	**1 (0)**	**1 (0)**
50%	II	0.51 (0.11)	0.59 (0.11)	0.62 (0.12)	0.71 (0.13)	0.52 (0.13)	**1 (0)**	**1 (0)**	**1 (0)**
foreground	III	0.63 (0.14)	0.70 (0.13)	0.77 (0.12)	0.91 (0.25)	0.59 (0.15)	**1 (0)**	**1 (0)**	**1 (0)**
	IV	0.42 (0.12)	0.49 (0.12)	0.53 (0.11)	0.53 (0.17)	0.43 (0.16)	0.92 (0.06)	0.92 (0.06)	**0.93 (0.05)**
Case 4:	I	0.36 (0.13)	0.38 (0.12)	0.40 (0.12)	0.29 (0.29)	0.29 (0.13)	0.79 (0.34)	0.95 (0.15)	**1 (0)**
25%	II	0.25 (0.10)	0.26 (0.09)	0.28 (0.09)	0.19 (0.08)	0.27 (0.09)	0.73 (0.36)	0.86 (0.28)	**0.99 (0.02)**
foreground	III	0.34 (0.09)	0.36 (0.09)	0.38 (0.09)	0.21 (0.14)	0.29 (0.10)	0.85 (0.29)	0.95 (0.17)	**1 (0)**
	IV	0.25 (0.08)	0.26 (0.07)	0.26 (0.07)	0.22 (0.13)	0.22 (0.11)	0.57 (0.28)	0.66 (0.25)	**0.73 (0.22)**

[1] Performance was measured by the proportions of true positives among the top T genes. T is the number of true positives in each simulated dataset. The mean and standard deviation of these proportions in the 50 simulated datasets were reported.

[*] There are four sub-cases in each of the simulated cases with the same amount of foreground columns.

Sub case I: no linear transformation, no cell-level noise;

Sub case II: only add linear transformation;

Sub case III: only add cell-level noise;

Sub case IV: add both linear transformation and cell-level noise.

[a] Query method using Pearson correlation coefficient.

[b] Query method using Spearman correlation coefficient.

[c] Query method using Kendall's τ.

[d] Query method using QDB.

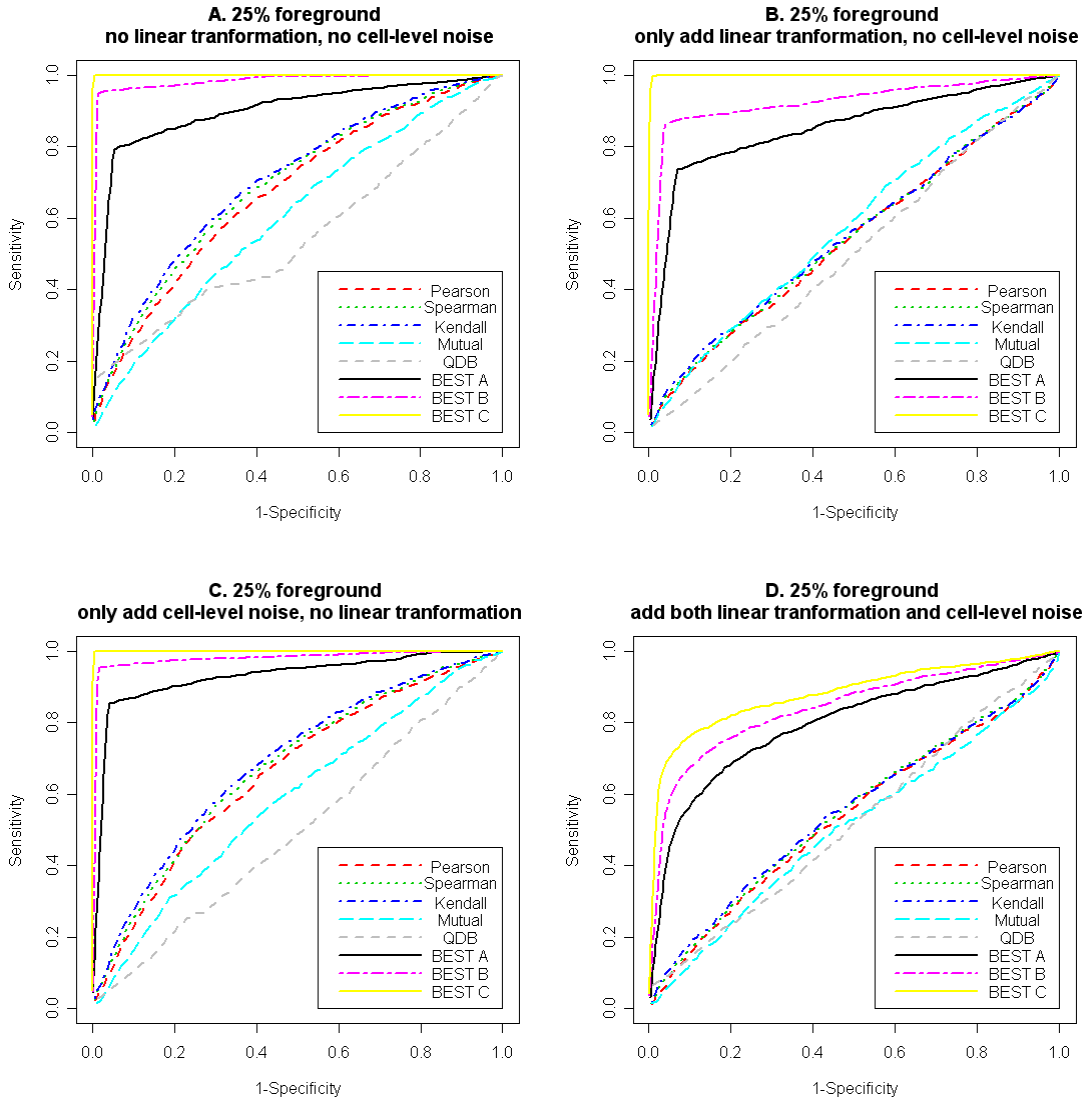
[e] Query method using mutual information.

[f] Query method using BEST.

[g] Query method using BEST allowing exclusion of individual cells from the foreground.

[h] Query method using BEST when fixing the indicator variables of five true target genes and five true experimental conditions as 1.

Figure 2.2 ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 25% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. A. No linear transformation nor cell-level noise. B. With linear transformation only. C. With cell-level noise only. D. With both linear transformation and cell-level noise.



From the simulation results, we see that all methods performed perfectly when all columns were foreground and no complicated correlation was present. In subsequent cases, the performances of all methods deteriorated with the inclusion of background columns, linear transformation and additional cell level noise. We observed that BEST is robust against added noise and complications and performed the best overall. Even in the most challenging case, in which the co-expression only occurred in 25% of the 50 columns, and half of the co-expressed genes were linearly transformed plus 10% additional cell-level noise, BEST still found 57% true co-expressed genes, and the AUC was 0.79. The simulation results also indicated that the version of BEST that allows cell-level noise has 5.4% to 8.2% higher AUCs compared to the version that does not consider cell-level noise. To evaluate the impact of incorporating existing knowledge into the model, we tested another version of BEST in which we fixed the indicator variables of five real target genes and five true foreground experimental conditions as 1. We found that in the most challenging case, AUCs further increased 1.0% to 7.6% compared to the version that considers cell-level noise. The superior performance of BEST in these synthetic datasets suggested that our algorithm worked well in the context of highly heterogeneous microarray data and was robust against moderately distorted data and sporadic outliers. Our model naturally accommodates existing biological knowledge which often results in further improvement in prediction accuracy. Among others, sophisticated methods such as QDB and the method based on mutual information performed better than the rest as expected. We acknowledge that our simulation scheme do not fit QDB well since it is a model-based bi-clustering algorithm not designed for the purpose of "querying *per se*".

2.3.2 Escherichia dataset

This dataset originally came from the study reported in (13). The authors conducted a comprehensive survey of gene expression profiles of all *E. coli* genes using 612 Affymetrix GeneChip arrays treated with 305 different experimental conditions. The goal of that study was to construct regulatory networks and determine the relative merits of different network inference algorithms on experimental data. RMA normalized data (39) was used in this study. This dataset consisted of 4,217 genes and 305 samples. We started with TF Leucine-responsive Regulatory Protein (Lrp) as the query gene. Faith et al. (2007) (13) listed Lrp as one of three TFs that show substantial connectivity in the network mapped by CLR and recommended it as an ideal test case. The *E. coli* Lrp is the best-studied member of the Lrp family, a global regulator in *E. coli* affecting the expression of many genes and operons (40). According to RegulonDB (41), Lrp has 61 experimentally verified transcription targets. We refer to the collection of these genes as the RegulonDB target set. Faith et al. (2007) (13) predicted potential transcription targets of Lrp using CLR, a mutual information-based algorithm. There were 43 genes predicted as Lrp targets at 60% precision and one gene was predicted as a Lrp target at 80% precision.

2.3.2.1 Query result from 100-gene test set

We tested BEST on this dataset to see if it could identify known target genes of Lrp. The 61 genes in the RegulonDB target set were included as positive genes. We also included 39 *E. coli* genes which displayed the most variation across the 305 experiments and not in the RegulonDB target set as negative genes. We used the 100-gene test set to compare performance of our algorithm with other query methods based on Pearson,

Spearman and Kendall correlation coefficients, mutual information and QDB. Using the default setting, BEST identified 28 target genes; 27 of them (96%) were in the RegulonDB target set (highly significant for enrichment with p-value of 1.27×10^{-6}). BEST also identified 143 experimental conditions (47%) as foreground. The log-likelihood trace plot suggested rapid convergence (Figure A2.4 in appendix). To compare the performance of our method with others, we plotted ROC curves (Figure 2.3). BEST achieved an AUC of 0.87, which was significantly higher than others (≤ 0.70). We also randomly selected 28 genes as targets for comparison, which showed an AUC of 0.52.

Figure 2.3 ROC curves for various query methods applying to the 100-gene test set selected from the *E. coli* microarray compendium. The area under the curves (AUC) are: Pearson correlation: 0.69; Spearman correlation: 0.69; Kendall's τ: 0.66; QDB: 0.70; Mutual information: 0.56; BEST: 0.87; Random control: 0.52.

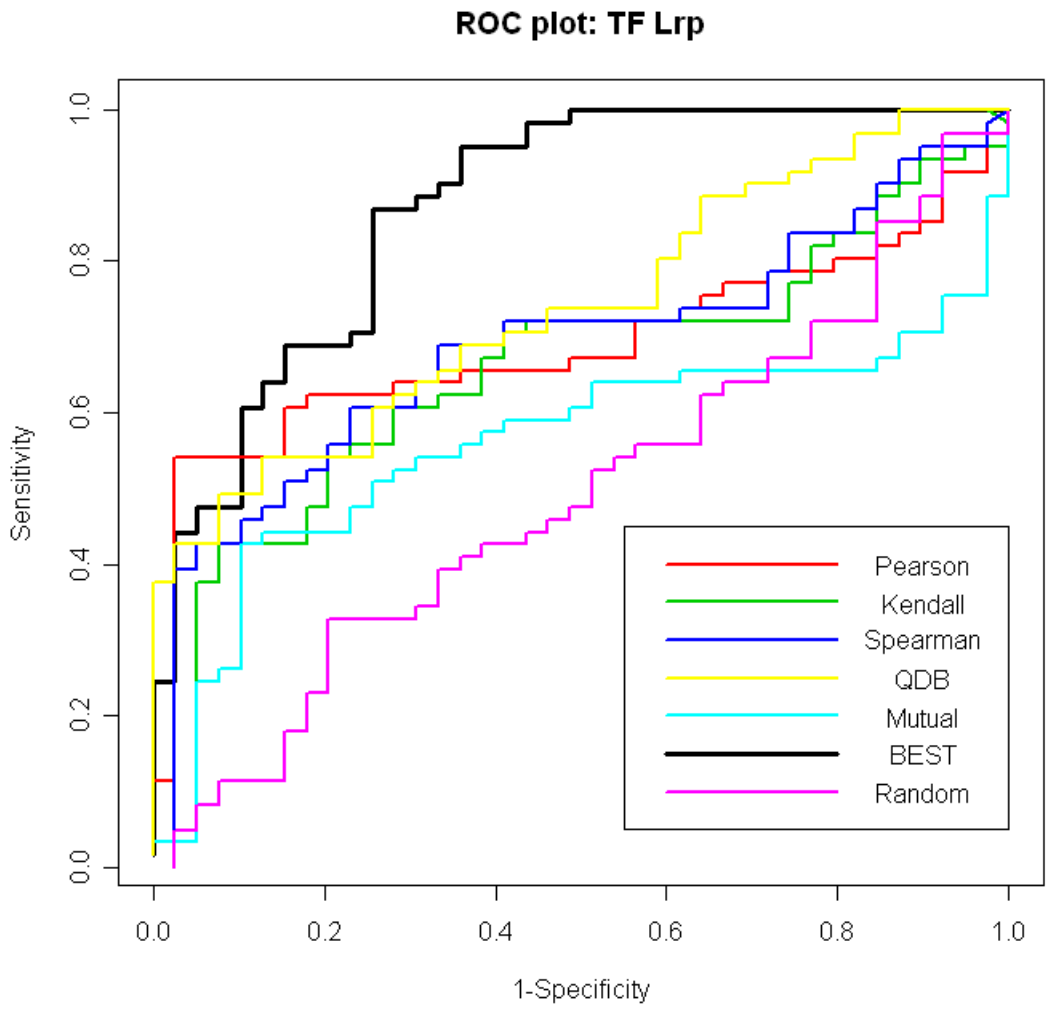
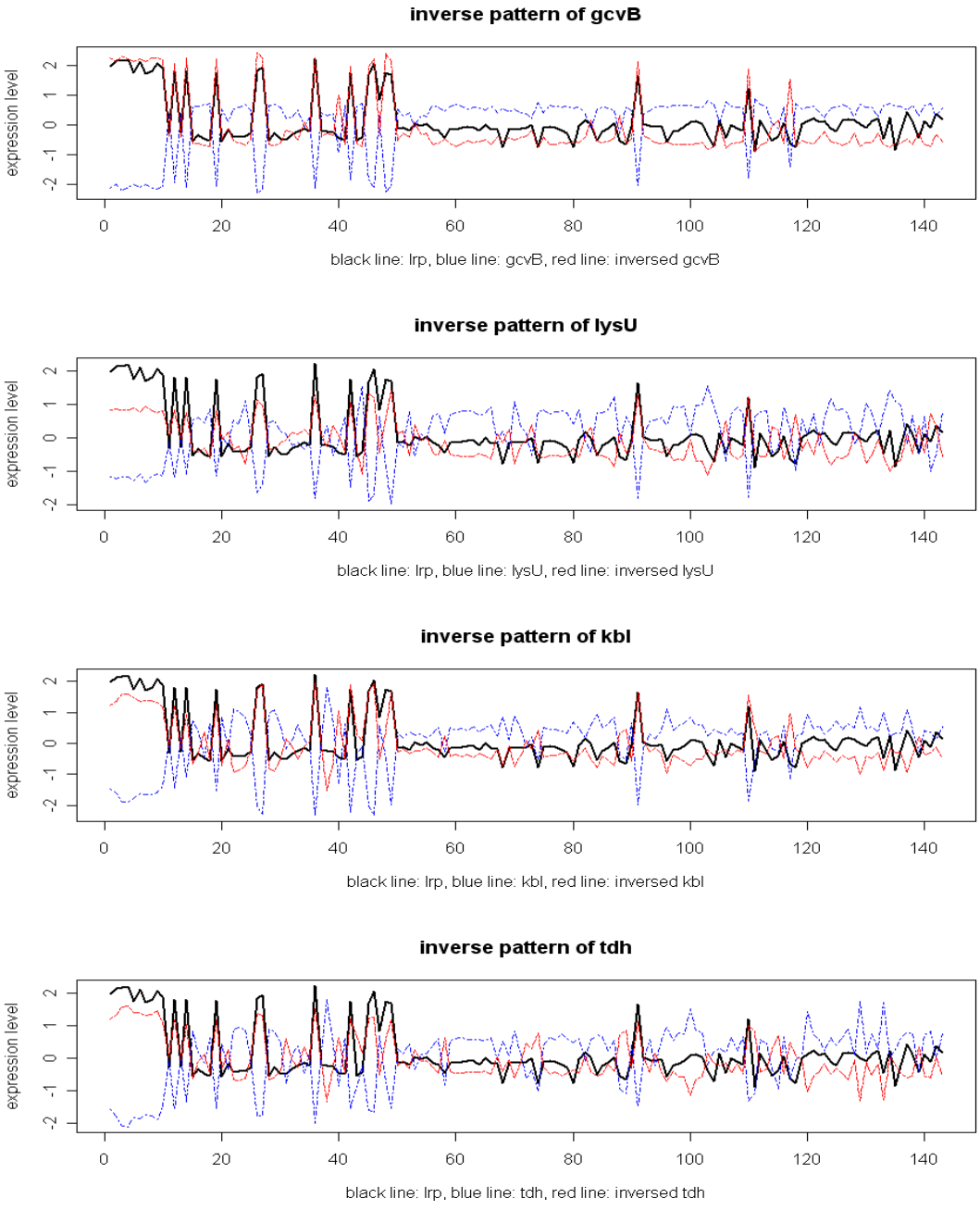


Figure 2.4 The original (blue line) and inverted (red line) expression profiles of gcvB, lysU, kbl and tdh compared to query gene Lrp. Black lines indicate the query gene—Lrp. Only the 143 foreground experimental conditions identified by BEST were shown in these plots. Results are from the 100-gene test set selected from the *E. coli* microarray compendium.



Among the 28 genes BEST identified (Table A2.2 in appendix), only one gene, gcvB, was not in the RegulonDB target set. gcvB is a regulatory RNA. It represses oppA, dppA, gltI and livJ expression and is regulated by gcvA and gcvR (42). Until now there has been no evidence to suggest gcvB is regulated by Lrp. However, the trace plot (Figure 2.4) showed that its expression profile, after inversion, is very close to the expression profile of Lrp. Its expression profile is also very close to that of three genes found in the RegulonDB target set, lysU, kbl and tdh (Table 2.2 and Figure 2.4). Furthermore, the scan of Lrp motif pattern (Figure A2.5 in appendix) indicates that there is a putative Lrp motif located in the intergenic region upstream of gcvB. Therefore we hypothesize that gcvB is also a target gene of Lrp (repressed by it).

Table 2.2 Information of the four genes showing inverse correlation patterns with Lrp identified by BEST when applied to the 100-gene test set selected from the *E. coli* microarray compendium. All but the first one, gcvB, are in the RegulonDB target set.

Rank	Gene name[a]	Log Bayes ratio	Positive/Negative[b]	RegulonDB[c]	CLR[d]	Motif distance[e]	Empirical p-value[f]
16	gcvB	107.80	Negative			414	0.0047
23	lysU	84.52	Negative	X		138	0.0044
24	kbl	81.47	Negative	X		33	0.0019
25	tdh	80.09	Negative	X			

[a] Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.

[c] "X" indicates that the predicted gene is in the RegulonDB target set.

[d] "X" indicates that the gene is predicted by CLR as a target gene.

[e] Motif distance is defined as the distance between the start position of the gene and the closest motif in the intergenic region upstream.

[f] Empirical p-value indicates the significance of conservation in the current motif, which is calculated as proportion of all possible motif locations in the complete *E. coli* genome that have likelihood ratios comparing between Lrp motif and background higher than that of the current motif.

Results from BEST also suggested that Lrp is likely to actively carry out most of its regulatory role under about half of all the 305 experimental conditions tested. To verify this hypothesis, we separately calculated Pearson correlation coefficients between the expression profiles of Lrp and genes in the RegulonDB target set in the 143 foreground conditions as well as the 162 background conditions. We found that the Pearson correlation coefficient in the foreground subset was indeed significantly higher than that of the background subset. A paired t-test comparing the two sets of correlation coefficients returned a p-value of 0.0079. When restricted to the 28 genes BEST identified as targets, the difference in the matched correlation coefficients became even more significant (p-value of 1.948×10^{-12}). Side-by-side box plots are shown in appendix Figure A2.6. The 305 experimental conditions were listed in Table A2.3 in appendix which was sorted by log Bayes ratio (larger values correspond to foreground). We found that many of the experimental conditions listed in the top portion are related to minimum media or stress which is consistent to what Faith et al. (2007) (13) found that including minimum media conditions will help identify Lrp targets.

The core of our model is the two-component Gaussian mixture for the expression levels obtained under the foreground experimental conditions. To verify this assumption, we plotted histograms of expression levels obtained under ten different experimental conditions, the top and bottom five when sorted by the Bayes ratio comparing whether an experimental condition is foreground or background. The histograms are shown in Figure A2.7 in appendix. As expected, we observed that the histograms from the top five

experimental conditions show strong bi-modal shapes while those from the bottom five do not.

2.3.2.2 Query result from 300-gene test set

To evaluate whether increased number of genes being queried and change in the proportion of negative genes affect BEST's performance, we added an additional 200 negative genes that showed high overall variations in all experiments to form a 300-gene test set.

Using the default setting in BEST, we identified 57 target genes (Table A2.4 in appendix) and 139 experimental conditions as foreground. Thirty-three of the target genes (58%) were in the RegulonDB target set (highly significant for enrichment with a p-value of 9.48×10^{-13}). A recent microarray analysis suggested that Lrp may affect transcription of as much as 10% of all *E. coli* genes (43). Therefore it is highly likely that many genes that are not in the RegulonDB target set are indeed regulated by Lrp. Trace plots of the 24 hypothetical Lrp target genes are shown in Figure A2.8 in appendix.

We next compared our result to the 43 genes CLR predicted as Lrp targets in (13). The 239 negative genes we selected actually contain four genes that are on the 43 CLR predicted target gene list but not in the RegulonDB target set. Three of them, metE, ompT and yagU were also identified by BEST as Lrp target genes. In fact, they ranked first, second and sixth in the 24 hypothetical Lrp targets genes listed in Table A2.4 in appendix. Interestingly, two of them, ompT and yagU have been confirmed to be bound by Lrp *in vivo* using ChIP-qPCR (13). Furthermore, the scan of Lrp motif indicates that all three genes contain a putative Lrp motif in their intergenic regions upstream.

We also plotted the histograms of expression levels obtained from the top and bottom five experimental conditions sorted by the Bayes ratio comparing whether an experimental condition is foreground or background (Figure A2.9 in appendix). We again observe strong bi-modal shapes in the histograms representing the top five experimental conditions but not in the histograms representing the bottom five.

2.3.2.3 Query result from other TFs

In addition to Lrp, we also ran BEST on six other TFs (PdhR, FecI, LexA, FlhC, FlhD and FliA) to test its performance. Among them, LexA, a major regulator of DNA repair, is known to have a single well-conserved DNA binding motif. It is one of the best-perturbed regulators in the microarray compendium due to the compendium's emphasis on DNA-damaging conditions (13). Other TFs either regulate a large number of genes or have substantial connectivity in the network mapped by the CLR Algorithm (13). For each TF, we built a test set including all its target genes listed in regulonDB, together with genes predicted by CLR as target. We also included ~100 genes which displayed the most variation across the 305 experimental conditions as negative signals. The complete results are summarized in the appendix and all BEST predicted target genes are listed in Tables A2.5 – A2.10 in appendix. From these lists, we see that except for PdhR, the majority of target genes listed in regulonDB were identified by BEST. For example, all six FecI target genes, 29 out of 30 FlhC target genes and 41 out of 42 FliA target genes were identified. Furthermore, BEST identified all CLR predicted target genes at 60% precision and numerous additional known target genes.

2.4 Discussion

In summary, we developed a model-based query algorithm based on the Bayesian model selection framework. BEST, a computer program implements this algorithm, is able to query large and heterogeneous microarray gene expression databases for regulon discovery. The query operation considered here can be viewed as a classification procedure where genes sharing similar expression profiles with the query gene belong to one group and the rest belong to the other. Therefore, we considered BEST a supervised learning tool. The key feature of BEST is its ability to recognize co-expression under only a subset of experimental conditions.

In microarray experiments with only a few sample/experimental conditions, the GBA principle has been successfully applied to identify regulons of key TFs (44). When the experimental conditions are abundant and heterogeneous such as in the case of microarray compendium, the previous strategy will not be as successful since most TFs are only active under certain specific conditions and beyond those conditions no tight correlation is expected between TF and its regulons. BEST is built under the hypothesis that the correlation between TF and its regulon only hold in a subset of conditions. The objective of BEST is to simultaneously predict regulon of a TF and the experimental conditions associated with them. Tests conducted on simulated as well as real datasets indicated that the new algorithm works well and outperforms methods based on global correlation measures, especially when there is substantial noise and moderate distortion in the data.

We are encouraged that when applying BEST to the real *E. coli* compendium data, the majority of genes predicted by BEST as Lrp targets are known target genes of

the TF. Interestingly, numerous genes identified show inversed correlation pattern with Lrp. Table 2.2 lists four such genes, three of them are known to be regulated by Lrp, and the other one showed a very similar pattern with the three known ones. None of these four genes is predicted by CLR. We also believe that many of the "false positive" genes are likely to be real Lrp target genes as well since as many as 10% of all *E. coli* genes are believed to be regulated by Lrp (43) which is significantly larger than the size of the current RegulonDB target set. We also tested major TFs whose target set is larger than ten. Querying these TFs showed that BEST is able to identify the majority of their known target genes. These results suggested that the hypothesis BEST assumed is reasonable. Using microarray compendium data, we are able to generate high confidence and testable hypothesis on TF-regulon relationships.

On the other hand, there are numerous genes in the RegulonDB target sets that were not identified by BEST. Visual inspection of these gene expression trace plots confirms that their expression profiles do not resemble the TF that is supposed to regulate them. This observation suggests that there are limitations on using the GBA principle on gene expression information alone to identify regulons of a TF. There are various reasons why GBA is insufficient to identify the full set of regulon. It is possible that the compendium does not include the experimental conditions under which these genes were regulated by the TF. It is possible that microarray gene expression data is not accurate enough due to measurement error and its limitation in quantifying low-level expression. It is also possible that due to the complexity in regulatory mechanism, some TF-regulon relationships do not imply co-expression under any condition. For example, the TF may require the presence of co-factors or signaling molecules to exert its regulatory function.

Other complex regulatory mechanisms such as post-translational modification, chromatin modification, and microRNA regulation may also explain what we observed.

In this study, we assumed that all columns are independent and there is no covariance. This is because replicates in our data have already been merged and adding covariance will significantly increase the complexity of our model. Admittedly, when there are biological or technical replicates, adding covariance in our model will improve the result. We plan to add this option in future releases of BEST.

It is possible to perform a genome-wide search using BEST for genes co-expressed with the query gene. To reduce computation time and to maximize the chance of finding biologically meaningful targets, we recommend a filtering step to reduce the search space. In this study, we adopted a variance filter, which is typical in large-scale gene expression clustering analysis (45) to remove genes that show less variation than the query gene when considering all experimental conditions. We tested this strategy on Lrp in *E. coli*. There are 524 genes (out of 4217 in total, 12%) with total expression variance greater than that of Lrp. They contain 30 genes (out of 61, 49%) that are in the RegulonDB target set. Running BEST with the default setting on this dataset identified 77 genes as targets. Among them, 18 are among the 30 known Lrp target genes (enrichment p-value is 3.32×10^{-9}). Compared to the CLR prediction in (13), seven of the 43 CLR predicted target genes that are not in the RegulonDB targets set are among the 524 genes tested. Six of them, gdhA, metE, ompT, pntA, thrA, yagU were also identified by BEST. All but metE have been confirmed *in vivo* as Lrp targets using ChIP-qPCR (13). The 139 experimental conditions identified by BEST as foreground are essentially the same as in the results from the 100- or 300-gene test sets. These results confirmed the

feasibility of our genome-wide search strategy. One can lower the variance threshold to expand the search space if longer computing time can be tolerated.

The statistical model adopted in BEST is closely related to those used in various model-based clustering methods designed for analyzing microarray data (46-51). However, as a supervised learning tool, BEST is able to automatically distinguish the two sets of genes using the expression profile of the pre-specified query gene. This is particularly valuable for searching specific expression patterns of interest. The user can even specify a custom expression pattern to search. In addition, our method allows linearly transformed expression patterns to be recognized and is robust against sporadic outliers in the data.

Our algorithm is built under the Bayesian model selection framework, which may easily incorporate prior biological information. For example, some genes or experimental conditions can be designated as targets or foreground. Similarly, informative priors on cell indicators can help to rule out some sporadic outliers.

MCMC-based methods are typically computation-intensive and therefore time-consuming. BEST's running time depends on the number of iterations and on the size of the dataset. In the study on *E. coli* microarray compendium dataset, using the default setting which is ten parallel chains each with 50 cycles, searching 100 genes takes about 30 minutes on a PowerMac with dual 2.5 GHz processors. Searching 300 and 524 genes takes about 3 hours and 30 hours respectively. A computer program named BEST has been developed based on the aforementioned algorithm. BEST can be downloaded at http://www.sph.umich.edu/csg/qin/BEST.

2.5 Appendix

2.5.1 Prior specification

We adopt standard conjugate priors for the model parameters. The prior parameters τ_{0j}, α_{0j}, β_{0j}, α_{1j} and β_{1j} are specified as follows:

First we estimate the linear transformation factor $\hat{a}_i$ using least square without intercept. To keep our model simple and avoid over-fitting, the linear transformation factor $\hat{a}_i$ is restricted to be significantly different from 0. The difference z_{ij} is defined as $y_{ij} - \hat{a}_i \cdot x_j$.

We calculate the sample mean $\hat{\mu}_{0j}$ and sample variance $\hat{v}_{0j}^2$ of the difference z_{ij} in each column, and set $\tau_{0j} = \hat{\mu}_{0j}$. Next we calculate the sample variance $\hat{\sigma}_0^2$ of column variance $\hat{v}_{0j}^2$, and choose α_{0j}, β_{0j}, such that the prior distribution $Inverse\ Gamma(\alpha_{0j}, \beta_{0j})$has mean $\hat{v}_{0j}$ and variance $\hat{\sigma}_0^2$.

To choose α_{1j}, β_{1j} in each column, we select z_{ij} with absolute value less than $\hat{\sigma}_{0j}/2$ in that column and calculate their sample variance $\hat{v}_{1j}^2$. Let $\hat{\sigma}_1^2$ be the variance of all $\hat{v}_{1j}^2$ s, and choose α_{1j} and β_{1j}, such that the prior distribution $Inverse\ Gamma(\alpha_{1j}, \beta_{1j})$ has mean $\hat{v}_{1j}$ and variance $\hat{\sigma}_1^2$.

We use non-informative priors for row indicator r_i, column indicator e_j and cell-level indicator c_{ij}, i.e.,$\pi_{r_i} = \pi_{e_j} = \pi_{c_{ij}} = 1/2$.

2.5.2 Full conditional distribution

The overall likelihood can be expressed as:

$$P(Z|R,E,\Theta) \propto \prod_{i=1}^{N}\prod_{j=1}^{M}\left(\frac{1}{\sigma_{1j}}\exp\left(-\frac{z_{ij}^2}{2\sigma_{1j}^2}\right)\right)^{r_i e_j}\left(\frac{1}{\sigma_{0j}}\exp\left(-\frac{(z_{ij}-\mu_{0j})^2}{2\sigma_{0j}^2}\right)\right)^{1-r_i e_j}$$

$$p(\sigma_{1j}^2) \sim Inverse\ Gamma(\alpha_{1j}, \beta_{1j}),$$

$$p(\sigma_{0j}^2) \sim Inverse\ Gamma(\alpha_{0j}, \beta_{0j}),$$

$$p(\mu_{0j}|\sigma_{0j}^2) \sim N(\tau_{0j}, \sigma_{0j}^2)$$

$$p(r_j) \sim Bernoulli(\pi_{r_i}),$$

$$p(e_j) \sim Bernoulli(\pi_{e_i}),$$

$$i = 1, \dots, N; j = 1, \dots, M$$

The marginal posterior distribution is:

$$p(R, E|Z) = \int P(R, E, \Theta|Z)\, d\Theta$$

$$\propto \int P(Z|R, E, \Theta) \prod_{i=1}^{N} p(r_i) \prod_{j=1}^{M} \left[p(\mu_{0j}|\sigma_{0j}^2) p(\sigma_{0j}^2) p(\sigma_{1j}^2) p(e_j)\right] d\Theta$$

After simplification, the log transformed marginal posterior distribution is:

$$\log p(R, E|Z) = -\sum_{j=1}^{M} \frac{1}{2} \log\left[1 + \sum_{i=1}^{N} (1 - r_i e_j)\right] + \sum_{j=1}^{M} \log\Gamma\left(\alpha_{0j} + \sum_{i=1}^{N} \frac{1 - r_i e_j}{2}\right)$$

$$+ \sum_{j=1}^{M} \log\Gamma\left(\alpha_{1j} + \frac{e_j}{2} \sum_{i=1}^{N} r_i\right) - \sum_{j=1}^{M} \left(\alpha_{1j} + \frac{e_j}{2} \sum_{i=1}^{N} r_i\right) \log\left(\beta_{1j} + \frac{e_j}{2} \sum_{i=1}^{N} z_{ij}^2 r_i\right)$$

$$- \sum_{j=1}^{M} \left(\alpha_{0j} + \sum_{i=1}^{N} \frac{1 - r_i e_j}{2}\right)$$

$$\times \log\left\{\frac{\tau_{0j}^2}{2} + \beta_{0j} + \sum_{i=1}^{N} \frac{z_{ij}^2 (1 - r_i e_j)}{2} - \frac{\left(\sum_{i=1}^{N} z_{ij}(1 - r_i e_j) + \tau_{0j}\right)^2}{2\left[1 + \sum_{i=1}^{N} (1 - r_i e_j)\right]}\right\}$$

$$+ \sum_{i=1}^{N} r_i \log\frac{\pi_{r_i}}{1 - \pi_{r_i}} + \sum_{j=1}^{M} e_j \log\frac{\pi_{e_j}}{1 - \pi_{e_j}} + constant$$

Allowing cell level noise, the overall likelihood is modified as follows:

$$P(Z|R,E,C,\Theta) \propto \prod_{i=1}^{N}\prod_{j=1}^{M}\left(\frac{\exp\left(-\frac{z_{ij}^2}{2\sigma_{1j}^2}\right)}{\sigma_{1j}}\right)^{r_i e_j c_{ij}}\left(\frac{\exp\left(-\frac{(z_{ij}-\mu_{0j})^2}{2\sigma_{0j}^2}\right)}{\sigma_{0j}}\right)^{1-r_i e_j c_{ij}}$$

$$p(\sigma_{1j}^2) \sim Inverse\ Gamma(\alpha_{1j}, \beta_{1j}),$$

$$p(\sigma_{0j}^2) \sim Inverse\ Gamma(\alpha_{0j}, \beta_{0j}),$$

$$p(\mu_{0j}|\sigma_{0j}^2) \sim N(\tau_{0j}, \sigma_{0j}^2)$$

$$p(r_j) \sim Bernoulli(\pi_{r_i}),$$

$$p(e_j) \sim Bernoulli(\pi_{e_i}),$$

$$p(c_{ij}) \sim Bernoulli\left(\pi_{c_{ij}}\right),$$

$$i = 1, \dots, N; j = 1, \dots, M$$

The marginal posterior distribution is:

$$p(R,E,C|Z) = \int P(R,E,C,\Theta|Z)d\Theta$$

$$\propto \int P(Z|R,E,C,\Theta)\prod_{i=1}^{N}p(r_i)\prod_{j=1}^{M}\left[p(\mu_{0j}|\sigma_{0j}^2)p(\sigma_{0j}^2)p(\sigma_{1j}^2)p(e_j)\right]\prod_{i=1}^{N}\prod_{j=1}^{M}p(c_{ij})\,d\Theta$$

After simplification, the log transformed marginal posterior distribution is:

$$\log p(R,E,C|Z) = -\sum_{j=1}^{M}\frac{1}{2}\log\left[1+\sum_{i=1}^{N}(1-r_i e_j c_{ij})\right] + \sum_{j=1}^{M}\log\Gamma\left(\alpha_{0j}+\sum_{i=1}^{N}\frac{1-r_i e_j c_{ij}}{2}\right)$$

$$+\sum_{j=1}^{M}\log\Gamma\left(\alpha_{1j}+\frac{e_j}{2}\sum_{i=1}^{N}r_i c_{ij}\right)-\sum_{j=1}^{M}\left(\alpha_{1j}+\frac{e_j}{2}\sum_{i=1}^{N}r_i\,c_{ij}\right)\log\left(\beta_{1j}+\frac{e_j}{2}\sum_{i=1}^{N}z_{ij}^2 r_i\,c_{ij}\right)$$

$$-\sum_{j=1}^{M}\left(\alpha_{0j}+\sum_{i=1}^{N}\frac{1-r_i e_j c_{ij}}{2}\right)$$

$$\times \log\left\{\frac{\tau_{0j}^2}{2} + \beta_{0j} + \sum_{i=1}^{N} \frac{z_{ij}^2(1 - r_i e_j c_{ij})}{2} - \frac{\left(\sum_{i=1}^{N} z_{ij}\left(1 - r_i e_j c_{ij}\right) + \tau_{0j}\right)^2}{2\left[1 + \sum_{i=1}^{N}\left(1 - r_i e_j c_{ij}\right)\right]}\right\}$$

$$+ \sum_{i=1}^{N} r_i \log\frac{\pi_{r_i}}{1 - \pi_{r_i}} + \sum_{j=1}^{M} e_j \log\frac{\pi_{e_j}}{1 - \pi_{e_j}} + \sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij} \log\frac{\pi_{c_{ij}}}{1 - \pi_{c_{ij}}} + constant$$

2.5.3 Detail protocol of microarray data analysis procedure using BEST

The *E. coli* dataset originally came from the study reported in Faith et al. (2007) (13). The authors conducted a comprehensive survey of gene expression profiles of all *E. coli* genes using 612 Affymetrix GeneChip arrays treated with 305 different experimental conditions. RMA normalized data (Faith et al., 2008) (39) was used in this study, which consisted of 4,217 genes and 305 samples. The detail of microarray data analysis procedure, such as microarray profiling, bacterial strains, steady-state experiments, time-course experiments, preparation of RNA and hybridization, external data, microarray normalization, are available at (Faith et al., 2007) (13).

Step 1: download microarray compendium data file "E_coli_v4_Builid_4_norm.tar.gz" from http://m3d.bu.edu/norm/?C=M;O=A. This zipped data file describes the normalized compendium dumps from M3D, which contains six files with expression data. "avg_E_coli_v4_Build_4_exps305probes4217.tab", the expression data file which contains 305 experimental conditions and 4,217 genes, was used in our study.

Step 2: get the expression profile of the query gene, for example: Lrp, from the microarray compendium "avg_E_coli_v4_Build_4_exps305probes4217.tab", which is the expression profile of Lrp across the 305 different experimental conditions.

Step 3: filter genes based on their variances. First, we calculated the variances of all 4,217 genes found in the microarray compendium. We then remove all genes whose variation across all experimental conditions is less than the query gene. This purpose of filtering is to reduce computation time and to maximize the chance of finding biological meaningful targets. For the query gene Lrp, there are 524 genes (out of 4,217 in total, 12%) with total expression variance greater than that of Lrp. We thus used these 524 candidate genes in our search.

Step 4: normalize the query gene and the 524 candidate genes. First, we calculated the mean and standard deviation across the 305 experimental conditions for each gene, and then normalize each of the gene expression levels by subtracting its mean and dividing by its standard deviation. After normalization, the query gene Lrp and the 524 candidate genes all have the same mean and variance (mean=0 and standard deviation=1).

Step 5: run BEST on the normalized gene expression levels using user-specified parameters such as the number of iteration in MCMC and the number of parallel chains.

Step 6: conduct motif search. We download position specific weight matrices (PSWM) from RegulonDB (http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt), and the complete *E. coli* genome from GenBank (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12_substr__MG1655/U00096.fna). We then calculated the log likelihood ratio comparing between the motif model and the background model on each possible start location in the intergenic regions

(up to 500 bp upstream) of genes identified by BEST. The locations with log likelihood ratio higher than a certain threshold are treated as putative motifs.

2.5.4 Query result from six other transcription factors

In addition to TF Leucine-responsive Regulatory Protein (Lrp), we used another six TFs (PdhR, FecI, LexA, FlhC, FlhD and FliA) as the query genes in this study to test BEST's performance. All six TFs have an almost equal number of target genes in ReglonDB and CLR prediction with an estimated 60% precision (Faith et al. 2007) (13).

2.5.4.1 Query result from PdhR

PdhR has five target genes in ReglonDB. CLR predicted four target genes with 60% precision. None of them is in the RegulonDB target set. We included these nine genes and another 91 negative genes to form a 100 gene test set. BEST found 27 target genes and 179 experimental conditions as foreground. Twenty-seven of BEST's target genes included two target genes in ReglonDB and four target genes predicted by CLR (p-value of 0.0110). We found three genes (uspE, cspD, aceA) with inversed pattern. Table A2.5 lists all PdhR target genes identified by BEST.

2.5.4.2 Query result from FecI

FecI has six target genes in ReglonDB. CLR predicted eight target genes with 60% precision. Eight of these nine predictions are not in the RegulonDB target set. We included these 13 genes and another 87 negative genes to form a 100 gene test set. BEST found 31 target genes and 169 experimental conditions as foreground. Thirty-one of BEST's target genes included all 13 target genes in ReglonDB and target genes predicted by CLR (p-value of 2.9×10^{-8}). We found no gene with inversed pattern. Table A2.6 lists all FecI target genes identified by BEST.

2.5.4.3 Query result from LexA

LexA has 16 target genes. CLR predicted 17 targets genes with 60% precision. 10 of these 17 predictions are not in the RegulonDB target set. We included these 26 genes and another 74 negative genes to form a 100 gene test set. BEST found 31 target genes and 237 experimental conditions as foreground. Thirty-one of BEST's target genes included 10 target genes in ReglonDB and all target genes predicted by CLR (p-value of 1.5×10^{-8}). We found one gene (uspE) with inversed pattern. Table A2.7 lists all LexA target genes identified by BEST.

2.5.4.4 Query result from FlhC

FlhC has 30 target genes in ReglonDB. CLR predicted 53 targets genes with 60% precision. 24 of these 53 predictions are not in the RegulonDB target set. We included these 54 genes and another 146 negative genes to form a 200 gene test set. BEST found 54 target genes and 266 experimental conditions as foreground. Fifty-four of BEST's target genes included 29 target genes in ReglonDB and all target predicted by CLR (p-value of 2.7×10^{-46}). We found no gene with inversed pattern. yjdA is the new hypothetical FlhC target gene identified by BEST in addition to false positive genes in Faith's prediction with 60% precision. Table A2.8 lists all FlhC target genes identified by BEST.

2.5.4.5 Query result from FlhD

FlhD has 46 target genes in ReglonDB. CLR predicted 46 target genes with 60% precision. Twenty of these 46 predictions are not in the RegulonDB target set. We included these 66 genes and another 134 negative genes to form a 200 gene test set. BEST found 55 target genes and 215 experimental conditions as foreground. Fifty-five of

BEST's target genes included 29 target genes in ReglonDB and all target genes predicted by CLR (p-value of 1.67×10^{-17}). We found two genes (micF, gadX) with inversed pattern. cheY, cheZ, flxA, micF, gadX and yjdA are the six new hypothetical FlhD target genes identified by BEST in addition to false positive genes predicted by CLR with 60% precision. Table A2.9 lists all FlhC target genes identified by BEST.

2.5.4.6 Query result from FliA

FliA has 42 target genes in ReglonDB. CLR predicted 56 target genes with 60% precision. Fifteen of these 56 predictions are not in the RegulonDB target set. We included these 57 genes and another 143 negative genes to form a 200 gene test set. BEST found 56 target genes and 281 experimental conditions as foreground. Fifty-six of BEST's target genes included 41 genes in ReglonDB and all target predicted by CLR (p-value of 4.08×10^{-47}). We found no genes with inversed pattern, and no new hypothetical FliA target gene identified by BEST in addition to false positive genes predicted by CLR with 60% precision. Table A2.10 list all FliA target genes identified by BEST.

2.5.5 Tables in appendix

Table A2.1. Performance comparison using area under the curve (AUC) of Receiver Operating Characteristic (ROC) among various methods for querying simulated microarray gene expression datasets. Best results are displayed in bold.

Case	Sub-case*	Pearson[a]	Spearman[b]	Kendall[c]	QDB[d]	Mutual[e]	BEST A[f]	BEST B[g]	BEST C[h]
Case 1:	I	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
100%	II	0.68	0.68	0.68	0.85	**1**	**1**	**1**	**1**
foreground	III	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	IV	0.66	0.71	0.71	0.79	0.85	**1**	**1**	**1**
Case 2:	I	0.98	**1**	**1**	**1**	0.98	**1**	**1**	**1**
75%	II	0.7	0.71	0.71	0.89	0.95	**1**	**1**	**1**
foreground	III	0.99	**1**	**1**	**1**	0.98	**1**	**1**	**1**
	IV	0.67	0.71	0.71	0.84	0.88	**1**	**1**	**1**
Case 3:	I	0.89	0.93	0.95	0.99	0.87	**1**	**1**	**1**
50%	II	0.67	0.69	0.69	0.89	0.81	**1**	**1**	**1**
foreground	III	0.88	0.92	0.94	0.92	0.84	**1**	**1**	**1**
	IV	0.59	0.62	0.64	0.77	0.69	0.98	**0.99**	**0.99**
Case 4:	I	0.68	0.69	0.71	0.55	0.61	0.91	0.98	**0.99**
25%	II	0.55	0.55	0.56	0.51	0.58	0.85	0.92	**0.99**
foreground	III	0.67	0.68	0.69	0.51	0.59	0.93	0.98	**0.99**
	IV	0.54	0.55	0.56	0.53	0.51	0.79	0.84	**0.88**

[*] There are four sub-cases in each of the simulated cases with the same amount of foreground columns.

Sub case I: no linear transformation;

Sub case II: only add linear transformation;

Sub case III: only add cell-level noise;

Sub case IV: add both linear transformation and cell-level noise.

[a] Query method using Pearson correlation coefficient.

[b] Query method using Spearman correlation coefficient.

[c] Query method using Kendall's τ.

[d] Query method using QDB.

[e] Query method using mutual information.

[f] Query method using BEST.

[g] Query method using BEST allowing exclusion of individual cells from the foreground.

[h] Query method using BEST when fixing the indicator variables of five true target genes and five true experimental conditions as 1.

Table A2.2 Information of the 28 potential Lrp target genes identified by BEST when applied to the 100-gene test set selected from the *E. coli* microarray compendium.

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	serA	131.81		X	X
2	leuA	129.99		X	X
3	leuL	128.72		X	X
4	gltD	128.22		X	X
5	leuD	123.19		X	X
6	ilvI	120.44		X	
7	ilvH	119.61		X	X
8	gltB	119.04		X	
9	leuC	116.22		X	X
10	livG	115.37		X	X
11	ilvE	114.30		X	
12	livK	113.30		X	X
13	leuB	110.28		X	X
14	livJ	109.72		X	
15	livM	108.48		X	
16	gcvB	107.80	negative		
17	serC	103.20		X	X
18	aroA	97.58		X	X
19	livH	94.76		X	X
20	livF	93.82		X	X
21	ilvL	90.43		X	
22	ilvD	89.88		X	
23	lysU	84.52	negative	X	
24	kbl	81.47	negative	X	
25	tdh	80.09	negative	X	
26	ilvG	79.47		X	
27	ilvM	71.23		X	X
28	ilvA	65.02		X	

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).
[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.
[c] BEST indentifies 27 genes among 61 target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.
[d] "X" indicates that the gene is predicted by CLR as a target gene.

Table A2.3 Information on the 305 distinct experimental conditions (among 612 different chips with replicates). The 305 experimental conditions are sorted by log Bayes ratio. BEST predicts the top 143 as foreground and the rest as background. Detail information of these 612 chips and 305 experimental conditions could be found in microarray compendium data file "E_coli_v4_Builid_4_norm.tar.gz" from the Gardner Lab. (http://m3d.bu.edu/norm/?C=M;O=A)

Rank	Experimental Conditions	Replicates	Log Bayes Ratio
1	M9_K_arcA_anaerobic	3	50.47
2	M9_WT	3	48.79
3	M9_K_appY_anaerobic	3	44.81
4	M9_K_soxS_anaerobic	3	43.21
5	M9_K_arcA	3	38.07
6	M9_K_oxyR_anaerobic	3	37.60
7	M9_K_soxS	3	37.23
8	M9_K_arcAfnr	3	37.02
9	M9_WT_anaerobic	4	35.75
10	M9_K_oxyR	3	35.52
11	lacZ_W1863_t0	1	35.24
12	M9_K_fnr	3	34.34
13	ccdB_MG1655_t30	2	34.14
14	M9_K_fnr_anaerobic	3	34.05
15	ccdB_W1872_t60	1	31.42
16	lacZ_W1863_t60	1	31.11
17	ccdB_W1872_t30	1	30.66
18	ccdB_W1872_t90	1	29.37
19	M9_K_appY	3	29.01
20	lacZ_MG1063_t0	2	27.61
21	ccdB_MG1655_t0	2	26.92
22	norfloxacin_chelator_MG1063_t0	1	26.75
23	ccdB_chelator_MG1063_t0	1	26.31
24	ccdB_chelator_MG1063_t60	1	26.12
25	lacZ_W1863_t30	1	25.47
26	fnr_K_fnrAnaerobic	4	25.21
27	cybr_N_log	2	24.95
28	MG1063_uninduced_t0	1	24.50
29	MG1063_uninduced_t60	1	24.36
30	MG1063_uninduced_t30	1	24.22
31	suspension_7hr	1	23.49

32	MG1655_ampicillin_t120	1	23.40
33	ccdB_chelator_W1872_t0	1	23.30
34	MG1655_ampicillin_t30	1	23.23
35	ik_H2_T3	1	23.04
36	fnr_wtAnaerobic	3	22.85
37	ik_L2_T3.5	1	22.80
38	norfloxacin_BW25113_t120	1	22.59
39	norfloxacin_BW25113recA_t60	1	22.39
40	har_S1_R_noIPTG	3	21.62
41	ccdB_chelator_MG1063_t30	1	21.60
42	cybr_O_log	2	21.56
43	biofilm_4hr	1	21.32
44	ph5.7_anaerobic	5	21.24
45	carbonSourceForaging	2	21.02
46	WT_MOPS_glucose	5	20.77
47	ik_L2_T4	1	20.38
48	M9_K_arcAfnr_anaerobic	3	20.28
49	WT_MOPS_glycerol	2	20.08
50	ik_H2_T3.5	1	19.86
51	MG1655_uninduced_t0	1	19.36
52	luc2_U_N0000	2	19.30
53	MG1655_ampicillin_t60	1	18.85
54	ccdB_BW25113recA_t120	1	18.84
55	ik_L2_T3	1	18.77
56	menC_U_N0075	3	18.76
57	crcB_U_N0075	3	18.69
58	MG1655_t480_aerobic	2	18.61
59	cpxR_U_N0075	3	18.31
60	era_U_N0075	3	18.26
61	crp_U_N0075	3	18.09
62	luc_U_N0075	3	18.00
63	gcvR_U_N0075	3	17.83
64	dnaA_U_N0075	3	17.79
65	menB_U_N0075	3	17.65
66	fis_U_N0075	3	17.62
67	ccdB_BW25113recA_t30	1	17.60
68	ccdB_MG1063_t0	2	17.18
69	ik_H2_T2.5	1	17.04
70	WT_N0075	2	17.04
71	ccdB_W1872_t0	1	16.89
72	MG1655_t720_aerobic	2	16.86

73	norfloxacin_BW25113recA_t30	1	16.59
74	MG1655_t270_anaerobic	2	16.35
75	lacZ_W1863_t90	1	16.27
76	rimI_U_N0075	3	16.19
77	minD_U_N0075	3	16.15
78	zipA_U_N0075	3	15.99
79	dinP_U_N0025	3	15.75
80	lacZ_MG1063_t30	2	15.74
81	mcrB_U_N0075	3	15.62
82	ccdB_BW25113recA_t180	1	15.53
83	MG1655_uninduced_t60	1	15.48
84	norfloxacin_chelator_MG1063_t0.1	1	15.34
85	yebF_U_N0075	3	15.21
86	MG1655_t150_aerobic	2	15.13
87	ccdB_BW25113_t180	1	15.01
88	biofilm_wt_noGlucose	1	14.83
89	MG1655_t225_anaerobic	2	14.75
90	ccdB_BW25113_t30	1	14.61
91	ast_pBADsup2	3	14.50
92	ph7	5	14.17
93	recA_U_N0025	3	13.88
94	MG1655_t300_aerobic	2	13.71
95	uspA_U_N0075	3	13.53
96	norfloxacin_MG1063_t0	1	13.47
97	mcrC_U_N0075	3	13.41
98	bcp_U_N0075	3	13.22
99	b2618_U_N0075	3	13.19
100	pepAA_t0	2	13.00
101	nupC_U_N0075	3	12.84
102	ldrA_U_N0075	3	12.34
103	ccdB_K12_t90	1	12.15
104	lacZ_K12_t60	1	12.09
105	BW25113_uninduced_t120	1	11.80
106	yoeB_U_N0075	3	11.71
107	minE_U_N0075	3	11.48
108	ph7_anaerobic	5	11.45
109	luc_U_N0025	3	11.44
110	ast_pBAD18	3	11.42
111	MG1655_t405_aerobic	2	11.28
112	sbcB_U_N0075	3	10.53
113	fklB_U_N0075	3	10.45

114	lacZ_MG1655_t0	1	10.36
115	MG1655_t1080_aerobic	2	9.96
116	gyrA_U_N0075	3	9.66
117	har_S0_R_noIPTG	5	9.52
118	norfloxacin_BW25113_t180	1	9.51
119	suspension_4hr	1	9.45
120	W3110_K_luxS	2	9.23
121	murI_U_N0075	3	8.80
122	MG1655_uninduced_t120	1	8.43
123	BW25113recA_uninduced_t180	1	7.97
124	folA_U_N0075	3	7.76
125	lacZ_K12_t30	1	7.69
126	cspF_U_N0075	3	7.61
127	dnaN_U_N0075	3	6.63
128	ccdB_BW25113recA_t0	1	6.57
129	W3110_wt	2	6.43
130	ccdB_K12_t30	1	5.91
131	hlpA_U_N0075	3	5.48
132	MG1655_uninduced_t30	1	4.20
133	biofilm_7hr	1	3.80
134	T60_N10000	3	3.16
135	MG1655_t180_anaerobic	2	2.78
136	WT_N0000	2	2.69
137	dinI_U_N0025	3	0.68
138	BW25113recA_uninduced_t120	1	0.29
139	ccdB_BW25113recA_t60	1	0.29
140	ik_L2_T4.5	1	0.22
141	dam_U_N0075	3	0.05
142	lacZ_MG1655_t60	1	-0.22
143	rstB_U_N0075	3	-0.71
144	pepAA_t30	2	-1.03
145	dnaT_U_N0075	3	-1.16
146	MGD1_t30	2	-1.26
147	norfloxacin_BW25113recA_t0	1	-1.62
148	gyrI_U_N0075	3	-1.83
149	MG1655_kanamycin_t120	1	-1.95
150	ik_H2_T8	1	-2.05
151	sulA_U_N0025	3	-2.29
152	umuD_U_N0025	3	-2.38
153	ccdB_K12_t60	1	-2.44
154	T0_N0000	3	-2.50

155	ik_H2_T4.5	1	-2.56
156	norfloxacin_BW25113_t30	1	-2.93
157	lacZ_MG1655_t90	1	-2.93
158	ik_H2_T6	1	-3.02
159	ik_L2_T2.5	1	-3.04
160	ruvA_U_N0025	3	-3.06
161	ccdB_BW25113_t0	1	-3.09
162	lexA_U_N0025	3	-3.54
163	ik_L2_T5.5	1	-3.60
164	holD_U_N0075	3	-3.87
165	nrdA_U_N0075	3	-4.06
166	ccdB_K12_t120	1	-4.15
167	WT_D_N0100	2	-4.29
168	pepCO_t30	2	-4.37
169	lon_U_N0025	3	-4.56
170	recA_D_N0100	2	-4.60
171	MG1655_kanamycin_t60	1	-4.62
172	MG1655_kanamycin_t30	1	-4.65
173	ccdB_BW25113_t60	1	-4.90
174	galF_U_N0075	3	-5.04
175	ruvC_U_N0075	3	-5.24
176	T48_N10000	3	-5.43
177	MG1063_uninduced_t120	1	-5.44
178	ccdB_chelator_W1872_t30	1	-5.66
179	lacZ_MG1063_t90	2	-5.84
180	BW25113recA_uninduced_t0	1	-5.95
181	uvrA_U_N0025	3	-6.13
182	IHF_U_N0075	2	-6.15
183	MOPS_K_dps_stationary2	1	-6.15
184	relA_U_N0025	3	-6.27
185	BW25113recA_uninduced_t30	1	-7.01
186	emrR_U_N0075	2	-7.21
187	ik_H2_T4	1	-7.22
188	cybr_O	2	-7.24
189	WT_MOPS_stationary3	2	-7.39
190	WT_MOPS_heatShock	1	-7.60
191	MOPS_K_dps_stationary	2	-7.61
192	cybr_N	2	-7.66
193	luc_U_N0000	3	-7.72
194	ph8.5_anaerobic	5	-7.86
195	pyrC_U_N0075	3	-8.00

196	ph5	5	-8.10
197	BW25113_uninduced_t30	1	-8.33
198	cybr_N_stat	2	-8.51
199	WT_MOPS_stationary2	2	-8.53
200	hscA_U_N0075	3	-8.60
201	ik_L2_T6	1	-8.67
202	ik_H2_T5.5	1	-8.94
203	mazF_U_N0025	3	-9.11
204	pET3d_t0	2	-9.19
205	luc2_U_N0025	2	-9.21
206	T24_N10000	3	-9.22
207	nrdB_U_N0075	2	-9.36
208	BW25113_uninduced_t180	1	-9.42
209	T36_N10000	3	-9.74
210	MG1655_spectinomycin_t30	1	-9.74
211	norfloxacin_BW25113_t60	1	-9.75
212	WT_MOPS_cipro2	1	-10.61
213	ccdB_MG1655_t60	2	-10.66
214	MGD1_t0	2	-10.98
215	WT_MOPS_acetate	2	-10.99
216	ccdB_chelator_MG1063_t120	1	-11.15
217	ccdB_MG1063_t30	2	-11.39
218	lacZ_K12_t120	1	-11.41
219	cybr_O_stat	2	-11.44
220	W3110_K_luxS_glucose	1	-11.44
221	lacZ_MG1063_t60	2	-11.71
222	ccdB_BW25113_t120	1	-11.87
223	ccdB_MG1655_t90	2	-11.91
224	ik_L2_T5	1	-12.14
225	WT_MOPS_proline	2	-12.22
226	recA_D_N0050	2	-12.34
227	WT_MOPS_stationary4	2	-12.42
228	WT_MOPS_cipro	1	-12.81
229	WT_N0025	2	-12.92
230	T12_N10000	3	-13.07
231	lacZ_MG1063_120	1	-13.07
232	BW25113recA_uninduced_t60	1	-13.11
233	fnr_K_fnrAerobic	3	-13.21
234	MG1655_t150_anaerobic	2	-13.42
235	norfloxacin_BW25113_t0	1	-13.55
236	recA_D_N0000	2	-13.59

237	WT_MOPS_acidShock	2	-13.77
238	MG1655_spectinomycin_t60	1	-13.82
239	MG1655_spectinomycin_t120	1	-14.22
240	biofilm_15hr	1	-14.39
241	K12_t360	3	-14.44
242	ccdB_K12_t0	1	-14.44
243	WT_MOPS_lateLog	3	-14.46
244	suspension_15hr	1	-15.01
245	W3110_wt_glucose	2	-15.23
246	har_S4_noIPTG	3	-15.91
247	norfloxacin_BW25113recA_t120	1	-16.73
248	ik_H2_T5	1	-16.96
249	har_S1_noIPTG	3	-16.97
250	norfloxacin_MG1063_t30	1	-17.00
251	lacZ_MG1655_t30	1	-17.38
252	pepCO_t0	2	-17.62
253	lacZ_K12_t0	1	-17.85
254	lacZ_K12_t90	1	-18.07
255	har_S1_IPTG	3	-18.16
256	MOPS_K_cspA	1	-18.20
257	MG1655_t86400_cecum	5	-18.47
258	BW25113_uninduced_t60	1	-18.67
259	norfloxacin_chelator_MG1063_t0.2	1	-18.95
260	norfloxacin_chelator_MG1063_t0.3	1	-19.00
261	MG1655_t1560_aerobic	2	-19.27
262	biofilm_24hr	1	-19.31
263	biofilm_K_yceP	1	-19.35
264	MG1655_norfloxacin_t30	1	-19.60
265	MG1655_norfloxacin_t120	1	-19.95
266	har_S0_noIPTG	3	-20.33
267	MOPS_K_dps	3	-20.40
268	cybr_KNO_N	2	-20.71
269	har_S4_IPTG	3	-21.49
270	ph8.7	5	-21.97
271	K12_t150_K_fis	3	-22.24
272	ik_L2_T8	1	-22.32
273	T24_N0000	3	-22.43
274	pET3d_t30	2	-22.91
275	K12_t90_K_fis	3	-23.30
276	BW25113_uninduced_t0	1	-23.45
277	K12_t360_K_fis	3	-23.71

278	biofilm_K_yceP_indole	2	-25.13
279	WT_N0050	2	-25.65
280	T60_N0000	3	-25.75
281	MOPS_K_hupB	1	-26.19
282	norfloxacin_MG1063_t60	1	-26.30
283	WT_MOPS_stationary	2	-26.59
284	K12_t240_K_fis	3	-26.70
285	K12_t150	3	-27.19
286	har_S4_R_IPTG	3	-27.85
287	norfloxacin_BW25113recA_t180	1	-29.03
288	ccdB_MG1063_t60	2	-29.70
289	MG1655_norfloxacin_t60	1	-30.08
290	ccdB_MG1063_t120	1	-31.01
291	ccdB_chelator_W1872_t60	1	-31.04
292	K12_t240	3	-32.33
293	MOPS_K_crp	3	-32.52
294	MG1063_uninduced_t180	1	-32.69
295	biofilm_K_tnaA	1	-32.85
296	MOPS_K_hns	3	-33.23
297	ccdB_chelator_W1872_t120	1	-33.23
298	suspension_24hr	1	-33.70
299	biofilm_wt_glucose	1	-34.55
300	har_S1_R_IPTG	3	-35.58
301	biofilm_K_trpE	1	-37.15
302	ccdB_MG1063_t90	2	-37.89
303	norfloxacin_MG1063_t120	1	-45.15
304	K12_t90	3	-45.89
305	har_S4_R_noIPTG	3	-51.28

Table A2.4 Information on the 57 potential Lrp target genes identified by BEST in the 300-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	serA	187.80		X	X
2	gltD	182.97		X	X
3	metE	180.95			X
4	leuL	178.03		X	X
5	leuD	175.01		X	X
6	leuA	174.43		X	X
7	gltB	173.30		X	
8	livG	172.80		X	X
9	livJ	172.44		X	
10	ilvE	172.35		X	
11	ompT †	169.81			X
12	pyrI	169.77			
13	livK	169.16		X	X
14	ilvH	168.41		X	X
15	leuC	168.10		X	X
16	ilvI	168.00		X	
17	gcvB	166.47	negative		
18	serC	164.09		X	X
19	livM	163.56		X	
20	leuB	163.18		X	X
21	pyrB	161.91			
22	yagU †	158.71			X
23	aroA	158.46		X	X
24	cysD	158.39			
25	ilvD	157.84		X	
26	lysU	157.56	negative	X	
27	livH	157.28		X	X
28	livF	155.74		X	X
29	stpA	153.87		X	
30	cysK	152.23			
31	pheL	151.32			
32	tnaC	149.80	negative		
33	dppA	148.70			
34	cysN	147.61			
35	kbl	145.76	negative	X	

36	treC	143.28	negative			
37	ilvL	142.44		X		
38	tdh	140.63	negative	X		
39	pyrL	140.4				
40	ilvC	139.63				
41	sdaA	138.77	negative	X		
42	sdaC	136.96	negative			
43	ilvA	136.74		X		
44	thrL	135.9				
45	hisL	135.55				
46	yeeD	133.82				
47	ilvM	131.85		X		X
48	treB	130.2	negative			
49	ompF	129.21		X		
50	fdoG	127.87	negative			
51	oppA	127.26		X		
52	oppB	124.59		X		
53	rmf	122.65				
54	oppF	122.24		X		
55	ynaJ	118.39				
56	ilvG	113.06		X		
57	sroF	109.54				

[a] Genes displayed here are sorted by the Log Bayes ratio (target gene versus non-target gene).

[b] Blank mean the target gene shows the same pattern as the query gene. Negative means the target gene shows the inversed pattern as the query gene.

[c] BEST indentifies 33 genes among 61 target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.

[d] "X" indicates that the gene is predicted by CLR as a target gene.

[†] Previously unknown targets of Lrp, experimentally verified by ChIP (Faith et al. 2007).

Table A2.5 Information on the 27 potential PdhR target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	recN	348.48			
2	intE	342.08			
3	recA	337.20			
4	tisB	336.43			
5	xisE	332.99			
6	araB	330.06			
7	araA	328.56			
8	sulA	328.12			
9	araD	327.70			
10	ymfJ	322.03			
11	ymfT	317.29			
12	ymfL	316.94			
13	araE	314.20			
14	murC	307.50			X
15	ftsW	306.20			X
16	murD	304.53			X
17	ndh	298.38			X
18	aceE	283.28		X	
19	aceF	279.67		X	
20	uspE	275.72	negative		
21	proV	274.19			
22	cspD	267.67	negative		
23	isrB	263.68			
24	spf	248.58			
25	cspA	239.93			
26	tisA	221.34			
27	aceA	76.68	negative		

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).

[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.

[c] BEST indentifies two genes among five target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.

[d] "X" indicates that the gene is predicted by CLR as a target gene.

Table A2.6 Information on the 31 potential FecI target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium.

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	ymfT	206.62			
2	ymfJ	200.23			
3	ymfL	196.67			
4	araD	189.08			
5	xisE	185.50			
6	araB	185.40			
7	araA	182.57			
8	recN	177.90			
9	araE	177.45			
10	tisB	146.18			
11	tisA	144.54			
12	sulA	133.96			
13	recA	118.89			
14	proV	81.52			
15	fecE	72.40		X	
16	fecB	67.88		X	
17	isrB	63.99			
18	fecD	63.63		X	
19	fecC	63.13		X	
20	fecA	62.00		X	
21	fhuF	54.12			X
22	ybaN	53.60			X
23	exbB	46.24			X
24	fhuA	44.64			X
25	exbD	43.74			X
26	fecR	43.29		X	X
27	bfd	33.30			
28	micF	32.89			
29	spf	29.53			
30	cspA	19.14			
31	entB	-0.34			X

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).

[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.
[c] BEST indentifies all six genes among six target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.
[d] “X” indicates that the gene is predicted by CLR as a target gene.

Table A2.7 Information on the 31 potential LexA target genes identified by BEST in the 100-gene test set extracted from the *E. coli* compendium

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	dinF	370.20		X	X
2	araB	365.72			
3	araA	365.42			
4	araE	359.80			
5	araD	358.89			
6	ymfJ	349.63			
7	ymfL	347.87			
8	recN	347.63		X	X
9	xisE	339.91			
10	yebG	333.06			X
11	ymfT	332.53			
12	dinI	320.63			X
13	recX	318.60			
14	umuD	316.82		X	X
15	tisB	314.57			
16	yafN	314.26			X
17	tisA	312.63			
18	dinD	312.63			X
19	uvrA	310.18		X	X
20	dinG	308.19			X
21	yafO	306.66			X
22	sulA	305.85		X	X
23	polB	299.75		X	
24	recA	292.55		X	X
25	umuC	289.20		X	X
26	dinB	282.91			
27	bssS	262.57			
28	ssb	262.39		X	
29	uvrD	259.91		X	
30	yebF	242.58			X
31	uspE	239.61	negative		

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).
[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.

[c] BEST indentifies ten genes among 16 target genes in RegulonDB. “X” indicates that the predicted gene is in the RegulonDB target set.
[d] “X” indicates that the gene is predicted by CLR as a target gene.

Table A2.8 Information on the 54 potential FlhC target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	flgE	422.41		X	X
2	fliA	417.80		X	X
3	flgC	415.00		X	X
4	flgB	412.97		X	X
5	flgG	406.95		X	X
6	flgH	406.29		X	X
7	flgD	403.29		X	X
8	flhD	401.18			X
9	motB	398.98			X
10	fliL	398.54		X	X
11	fliN	398.03		X	X
12	flgI	397.19		X	X
13	flgK	396.32			X
14	flgA	396.03		X	X
15	fliK	389.85		X	X
16	fliM	388.79		X	X
17	flgF	388.45		X	X
18	motA	387.71			X
19	cheA	386.19			X
20	cheW	385.20			X
21	fliZ	384.50		X	X
22	fliJ	384.25		X	X
23	flgM	382.36			X
24	fliF	382.00		X	X
25	flgN	380.73			X
26	fliS	380.29			X
27	cheY	378.15			X
28	flgJ	372.04		X	X
29	cheZ	371.92			X
30	cheR	371.27			X
31	yecR	370.70			X
32	cheB	369.11			X
33	fliG	367.55		X	X
34	fliC	366.59			X
35	flgL	366.38			X
36	fliH	362.55		X	X

37	fliD	360.82		X
38	fliP	356.28	X	X
39	fliQ	350.16	X	X
40	tar	347.51		X
41	fliI	345.24	X	X
42	ycgR	344.20		X
43	tap	338.41		X
44	fliE	326.37	X	X
45	flxA	325.67		X
46	fliO	324.78	X	X
47	ymdA	314.15		X
48	flhA	301.22	X	X
49	flhE	300.16	X	X
50	flhB	290.68	X	X
51	fliR	275.89	X	X
52	yhjH	272.24		X
53	tsr	255.05		X
54	yjdA	206.14		

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).
[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.
[c] BEST indentifies 29 genes among 30 target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.
[d] "X" indicates that the gene is predicted by CLR as a target gene.

Table A2.9 Information on the 67 potential FlhD target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	flhC	333.41			X
2	flgE	312.18		X	X
3	flgB	311.50		X	X
4	flgH	311.46		X	X
5	fliA	311.18		X	X
6	flgC	310.34		X	X
7	flgG	307.33		X	X
8	flgD	299.71		X	X
9	flgK	299.55			X
10	flgI	299.12		X	X
11	fliZ	298.07		X	X
12	flgA	296.58		X	X
13	motB	294.72			X
14	cheW	293.40			X
15	flgF	293.22		X	X
16	fliN	292.68		X	X
17	fliL	292.29		X	X
18	flgN	291.93			X
19	fliF	290.99		X	X
20	flgM	290.12			X
21	fliM	288.74		X	X
22	cheA	288.54			X
23	fliK	285.84		X	X
24	fliS	282.94			X
25	motA	282.93			X
26	yecR	282.39			X
27	flgL	281.34			X
28	fliJ	279.79		X	X
29	fliC	275.71			X
30	flgJ	275.32		X	X
31	fliD	271.73			X
32	cheR	271.62			X
33	fliP	270.94		X	X
34	fliG	270.80		X	X
35	cheB	269.55			X
36	cheY	269.30			

37	fliH	264.98		X	
38	tar	264.64			X
39	cheZ	264.38			
40	fliE	260.07		X	X
41	fliI	256.40		X	X
42	ycgR	251.77			X
43	flxA	251.29			
44	fliQ	251.26		X	X
45	flhE	231.94		X	
46	fliO	229.43		X	X
47	flhB	222.84		X	
48	flhA	221.21		X	X
49	ymdA	218.56			X
50	fliR	196.03		X	X
51	tsr	192.51			X
52	yibT	191.41			
53	yhjH	191.29			X
54	yjbJ	188.62			
55	hdeB	187.86			
56	slp	184.57			
57	ompF	181.05			
58	micF	178.49	negative		
59	gadE	178.32	negative		
60	hdeA	177.08			
61	hdeD	176.26			
62	gadX	173.90			
63	gadB	171.77			
64	gadA	165.48			
65	yjdA	148.57			
66	bssS	122.36			
67	ygiW	109.00			

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).
[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.
[c] BEST indentifies 29 genes among 46 target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.
[d] "X" indicates that the gene is predicted by CLR as a target gene.

Table A2.10 Information on the 56 potential FliA target genes identified by BEST in the 200-gene test set extracted from the *E. coli* compendium

Rank	Gene Name[a]	Log Bayes Ratio	positive/negative[b]	RegulonDB[c]	CLR[d]
1	fliZ	524.28		X	X
2	flgE	520.41		X	X
3	flgC	518.79		X	X
4	flgB	510.76		X	X
5	flgG	506.72		X	X
6	flgD	506.65		X	X
7	flgN	503.87			X
8	flgK	503.51			X
9	flgH	490.47		X	X
10	flgM	488.91			X
11	fliD	484.75		X	X
12	cheW	484.68		X	X
13	cheA	480.61		X	X
14	motB	470.69		X	X
15	motA	469.88		X	X
16	flgL	469.11		X	X
17	flgA	467.86			X
18	fliK	463.84		X	X
19	fliS	459.14		X	X
20	fliN	456.92		X	X
21	flgF	456.74		X	X
22	cheZ	453.19		X	X
23	cheR	452.47		X	X
24	fliL	452.27		X	X
25	fliJ	452.03		X	X
26	flgI	448.03		X	X
27	cheB	443.04		X	X
28	tar	439.80			X
29	fliC	439.60			X
30	fliF	435.47		X	X
31	fliM	431.91		X	X
32	fliG	429.41		X	X
33	cheY	425.70		X	X
34	flgJ	423.63		X	X
35	fliP	420.70		X	X
36	yecR	418.55			X

37	ycgR	415.26		X
38	tap	414.20	X	X
39	fliQ	404.20	X	X
40	fliH	389.79	X	X
41	flxA	377.76	X	X
42	fliT	369.77	X	X
43	ymdA	364.62		X
44	fliO	361.51	X	X
45	fliI	355.28	X	X
46	fliE	354.40	X	X
47	flhC	348.31		X
48	flhB	335.53	X	X
49	flhE	321.02	X	X
50	flhA	317.85	X	X
51	fliR	312.35	X	X
52	yhjH	310.92		X
53	flhD	305.93		X
54	tsr	281.05		X
55	ves	233.09		X
56	yjdA	187.37		X

[a] Genes displayed here are sorted by Log Bayes ratio (target gene versus non-target gene).
[b] Blank indicates that the target gene shows the same pattern as the query gene. Negative indicates that the target gene shows the inversed pattern as the query gene.
[c] BEST indentifies 41 among 42 target genes in RegulonDB. "X" indicates that the predicted gene is in the RegulonDB target set.
[d] "X" indicates that the gene is predicted by CLR as a target gene.

2.5.6 Figures in appendix

Figure A2.1 ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 100% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. A. No linear transformation nor cell-level noise. B. With linear transformation only. C. With cell-level noise only. D. With both linear transformation and cell-level noise.

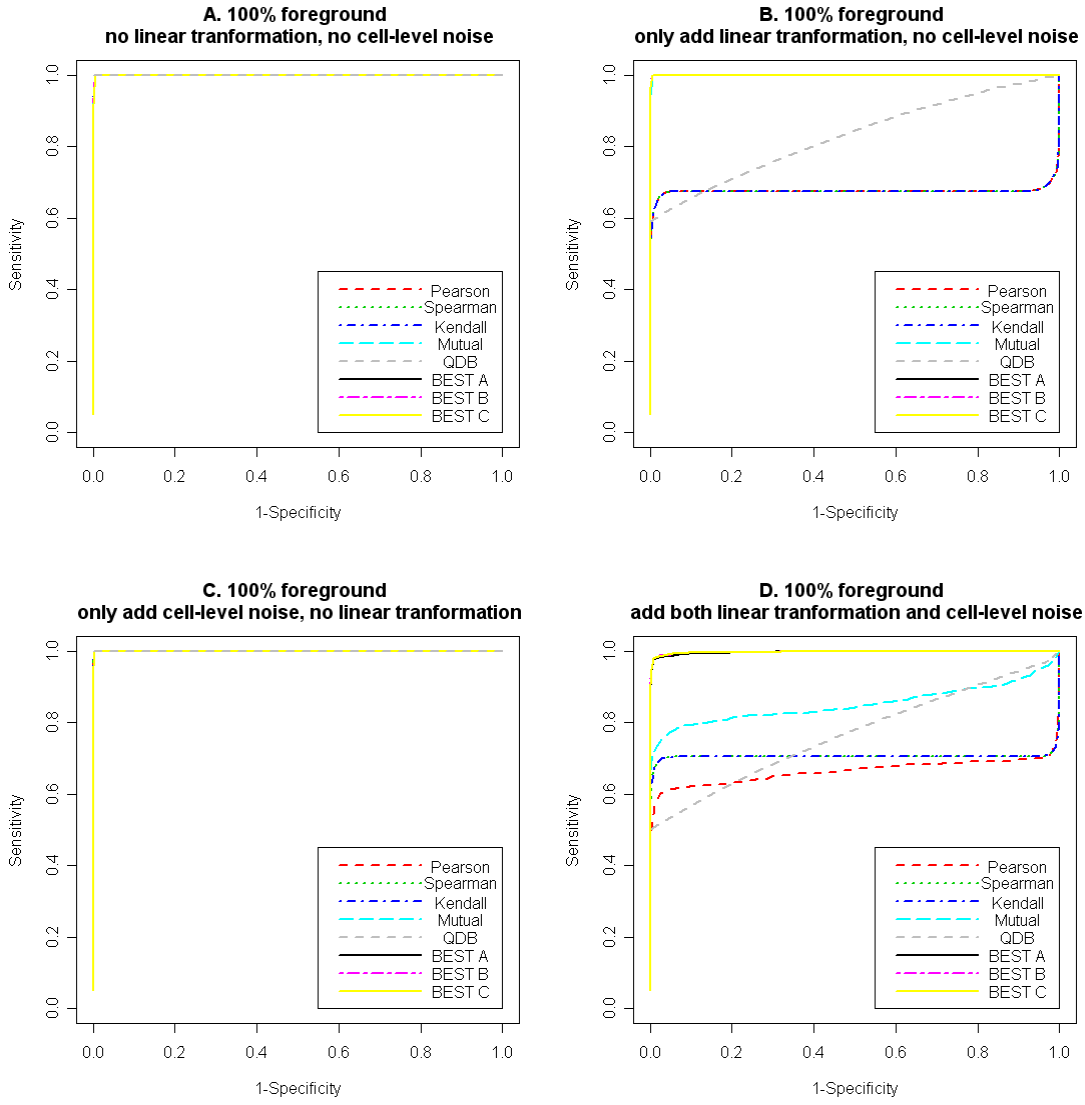


Figure A2.2 ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 75% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. A. No linear transformation nor cell-level noise. B. With linear transformation only. C. With cell-level noise only. D. With both linear transformation and cell-level noise.

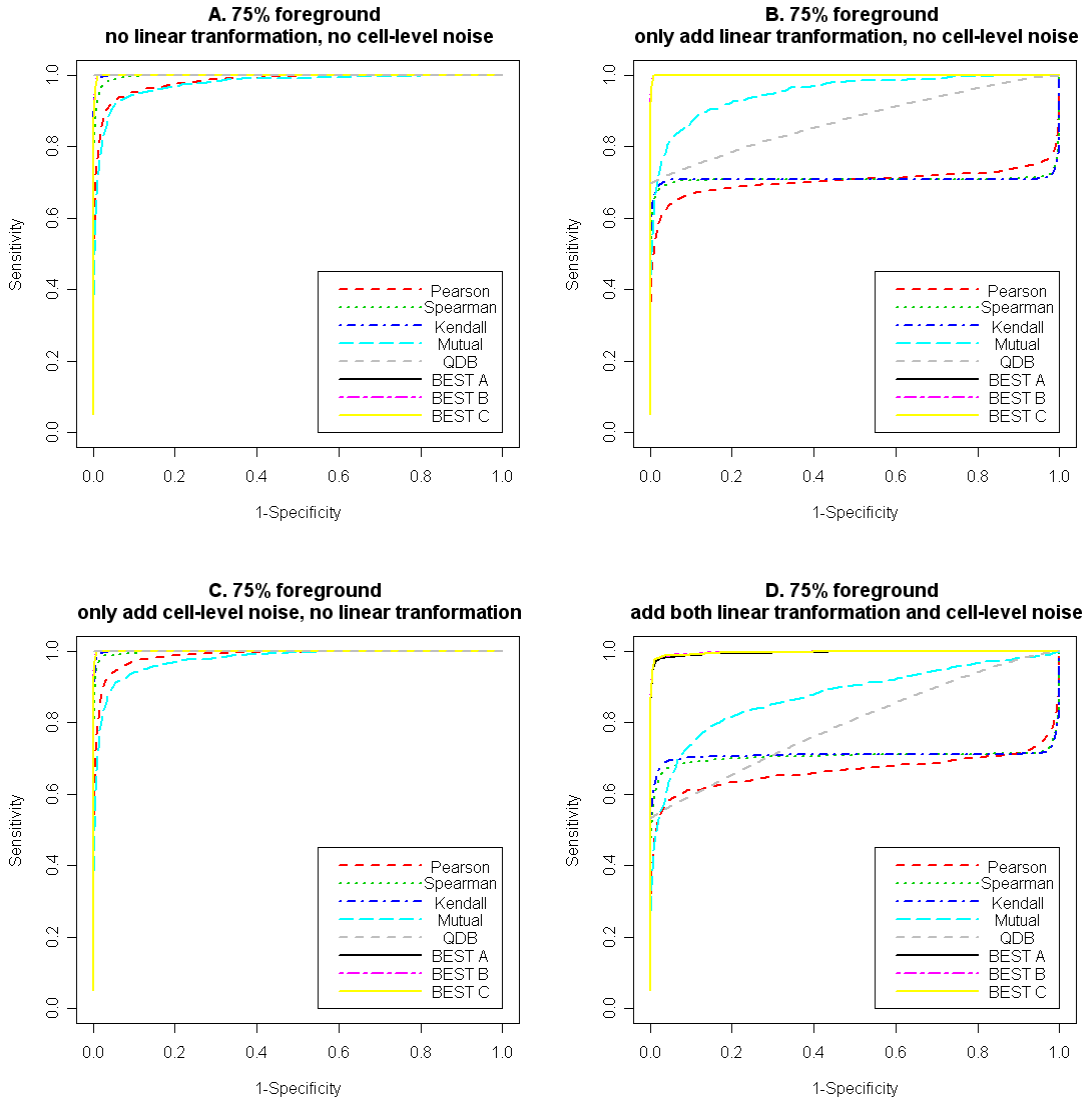


Figure A2.3 ROC curves for various query methods when applying to synthetic datasets simulated under different settings and when there are 50% foreground columns. BEST A default setting; BEST B allowing exclusion of individual cells from the foreground; BEST C fixing the indicator variables of five true target genes and five true experimental conditions as 1. A. No linear transformation nor cell-level noise. B. With linear transformation only. C. With cell-level noise only. D. With both linear transformation and cell-level noise.

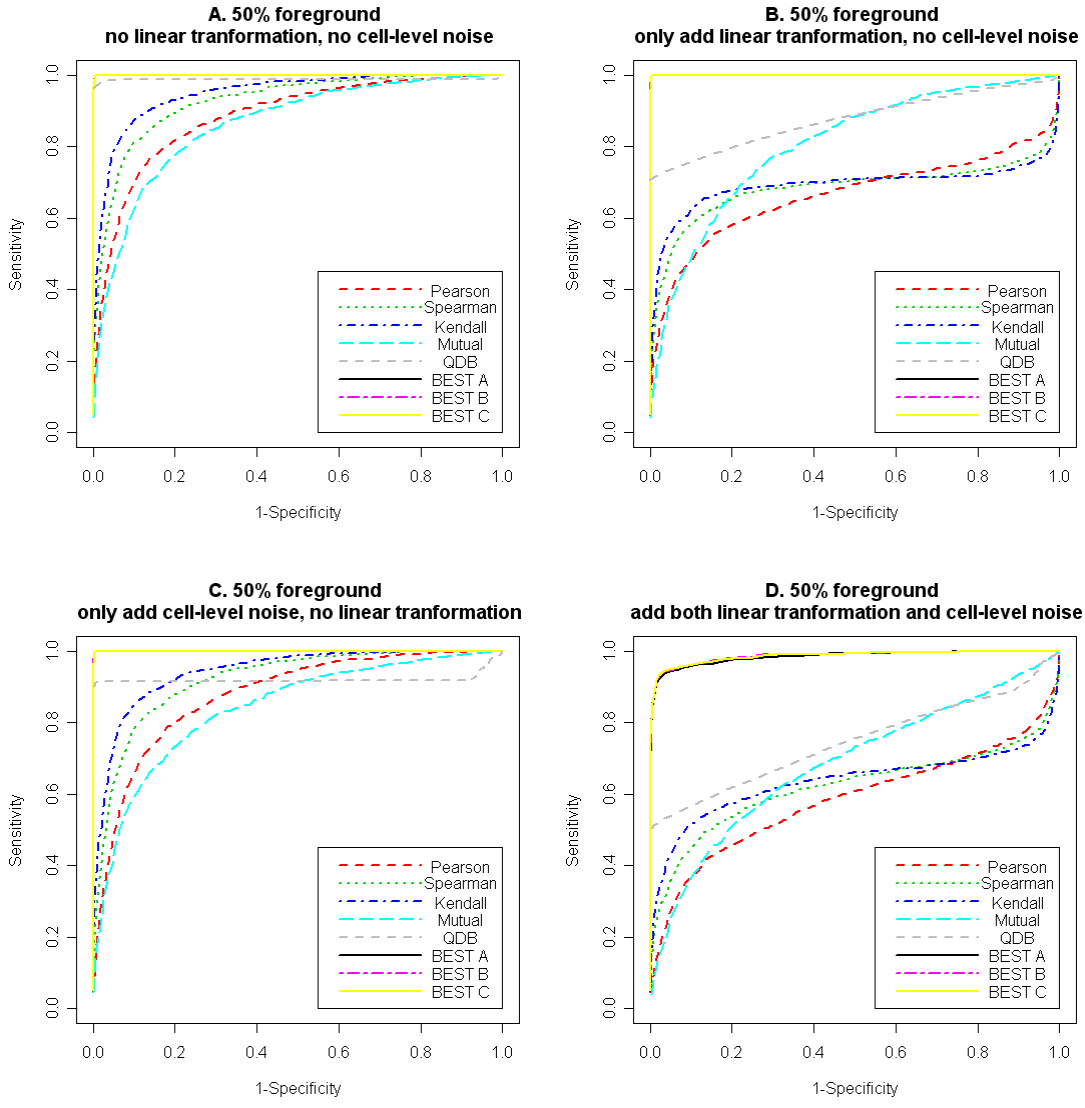


Figure A2.4 Log-likelihood trace plots of the ten parallel chains resulted from the BEST run on 100-gene and 300-gene test sets selected from the *E. coli* microarray compendium.

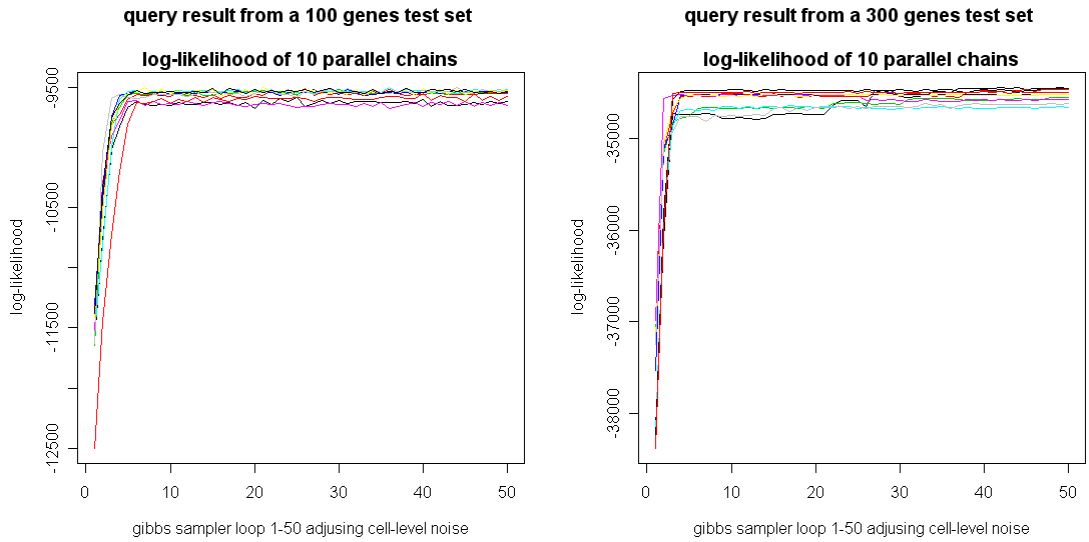
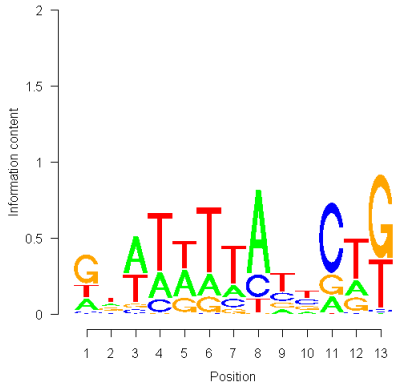


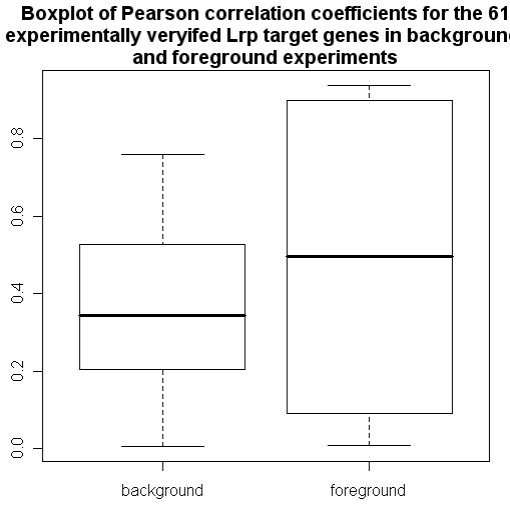
Figure A2.5 Sequence logo plot (52) and position specific weight matrix (PSWM) for the motif of transcription factor Lrp. Lrp motif is downloaded from regulonDB: http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt. The logo plot was generated by the seqLogo program (53).



Position	1	2	3	4	5	6	7	8	9	10	11	12	13
A	11	14	27	14	21	11	10	37	6	6	8	12	0
C	2	7	3	7	0	0	7	10	12	15	35	1	1
G	28	13	3	0	11	8	4	0	8	11	10	10	33
T	12	19	20	32	21	34	32	6	27	21	0	30	19

Figure A2.6 Boxplots of Pearson correlation coefficients. A. Boxplots of Pearson correlations between expression profiles of the 61 experimentally verified Lrp target genes and Lrp. The left one summarize correlations measured in the 162 background experiments and the right one summarize correlations measured in the 143 foreground experiments. A paired t-test comparing the two sets of correlation coefficients returns a p-value of 0.0079. B. Boxplots of Pearson correlations between expression profiles of the 28 genes BEST indentified as Lrp target. The left one summarize correlations measured in the 162 background experiments and the right one summarize correlations measured in the 143 foreground experiments. A paired t-test comparing the two sets of correlation coefficients returns a p-value of 1.948×10^{-12}.

A.



B.

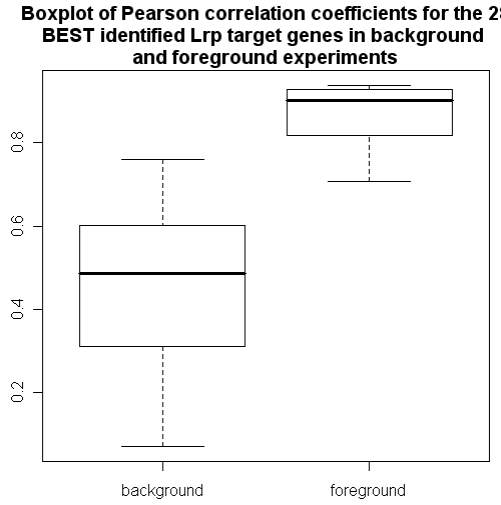
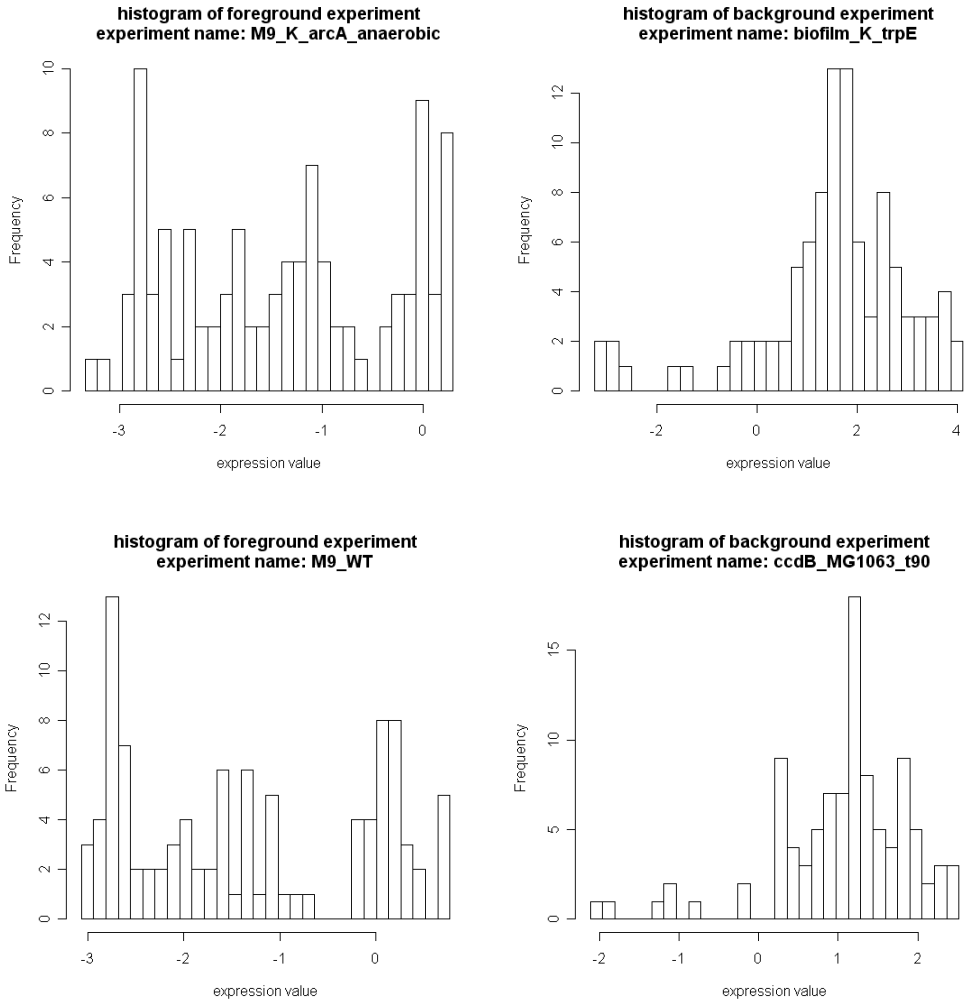
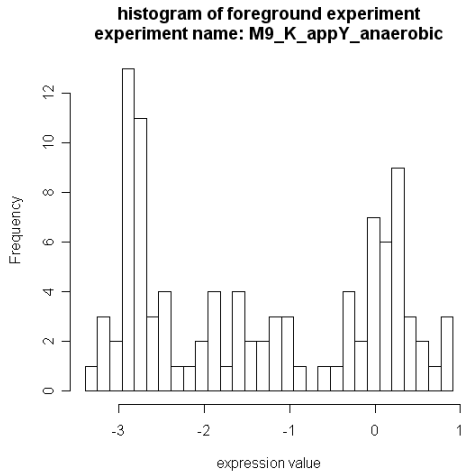


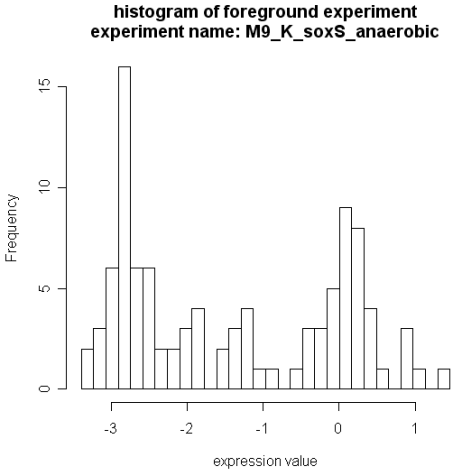
Figure A2.7 A. Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). B. Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 100-gene test set selected from the *E. coli* microarray compendium.

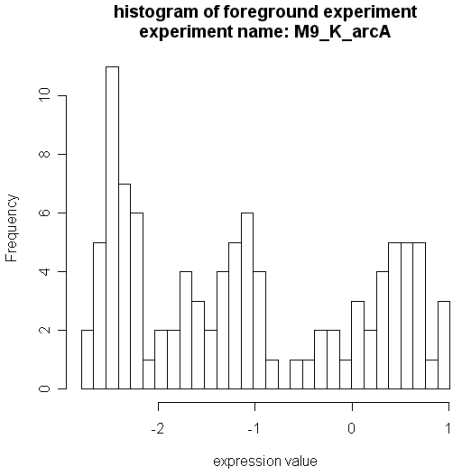
A. B.



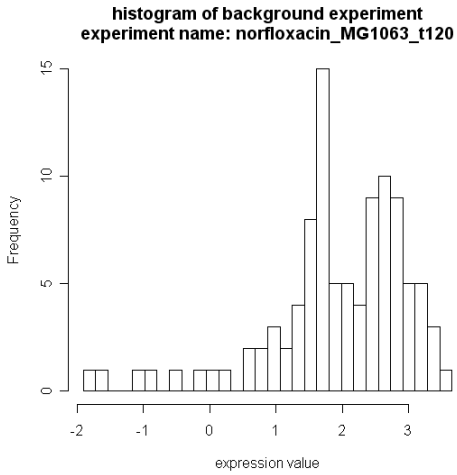
A.

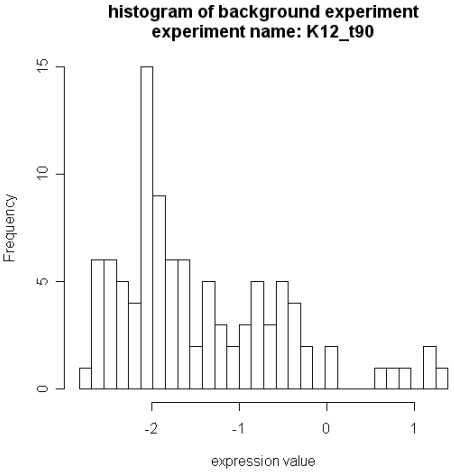

histogram of foreground experiment
experiment name: M9_K_appY_anaerobic
Frequency
0
2
4
6
8
10
12
-3
-2
-1
0
1
expression value


histogram of foreground experiment
experiment name: M9_K_soxS_anaerobic
Frequency
0
5
10
15
-3
-2
-1
0
1
expression value


histogram of foreground experiment
experiment name: M9_K_arcA
Frequency
0
2
4
6
8
10
-2
-1
0
1
expression value

B.


histogram of background experiment
experiment name: norfloxacin_MG1063_t120
Frequency
0
5
10
15
-2
-1
0
1
2
3
expression value


histogram of background experiment
experiment name: K12_t90
Frequency
0
5
10
15
-2
-1
0
1
expression value

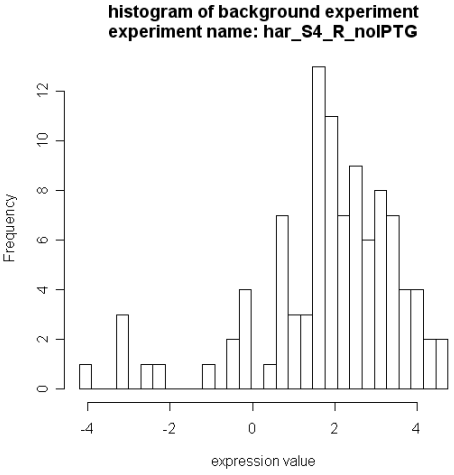
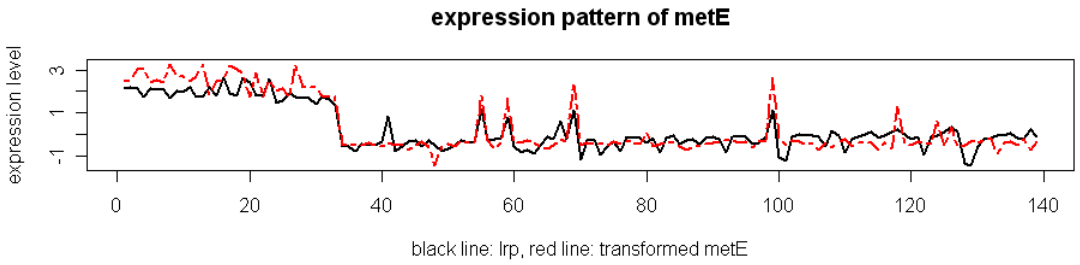
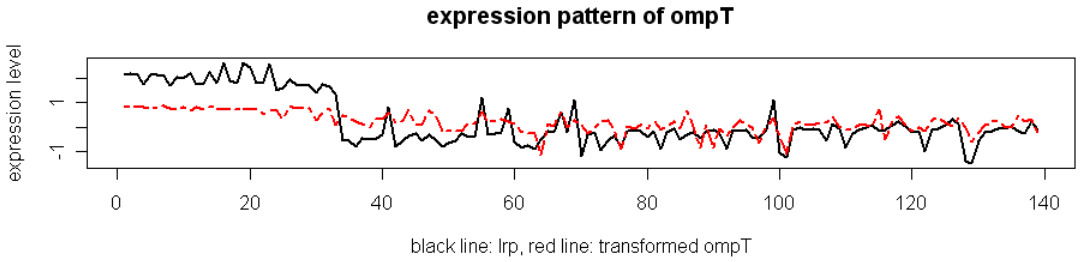
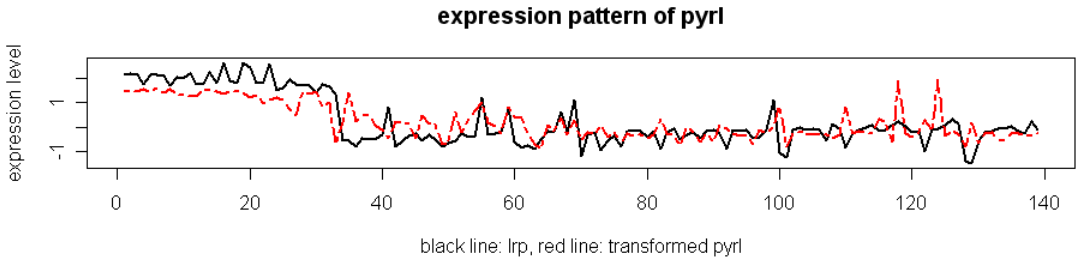
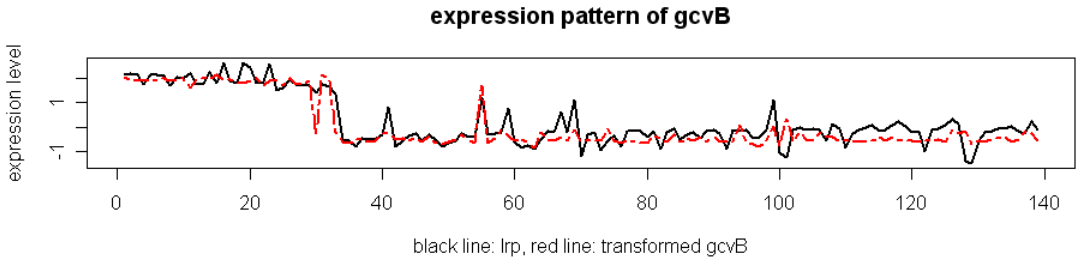

histogram of background experiment
experiment name: har_S4_R_noIPTG
Frequency
0
2
4
6
8
10
12
-4
-2
0
2
4
expression value

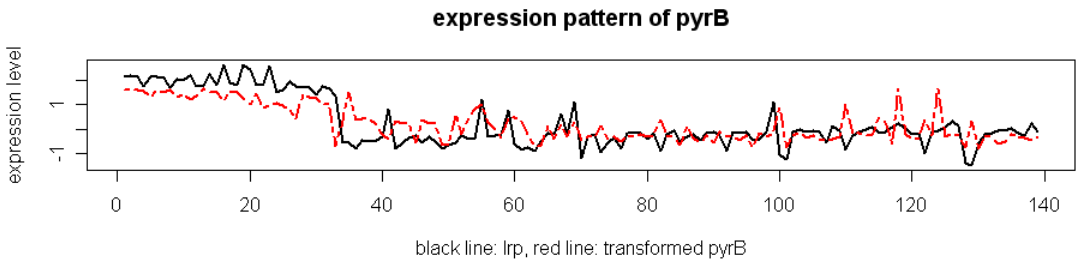
Figure A2.8 Trace plots of 24 predicted Lrp target genes identified by BEST that are not in the RegulonDB target set. Black lines indicate the query gene—Lrp, the red line indicate the potential target genes. Only the 139 foreground experimental conditions identified by BEST were shown in these plots.

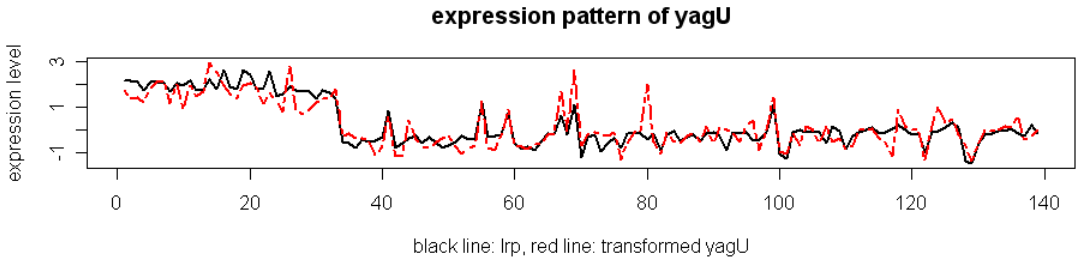


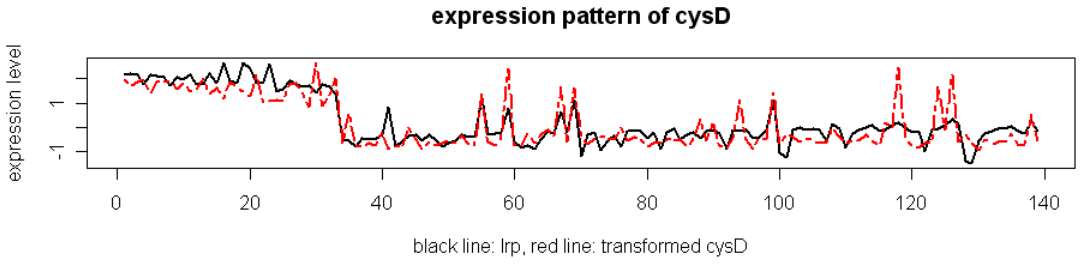


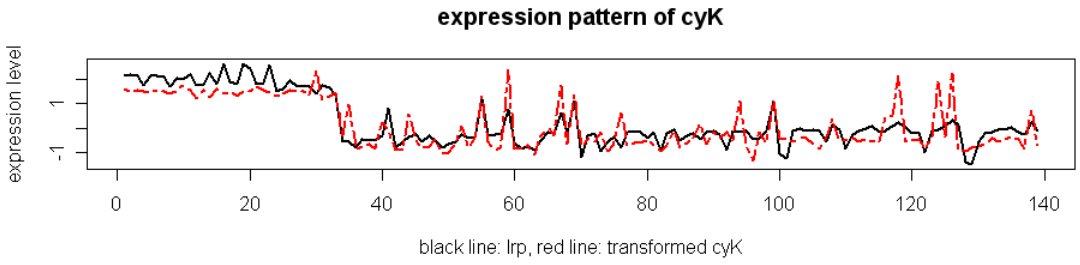


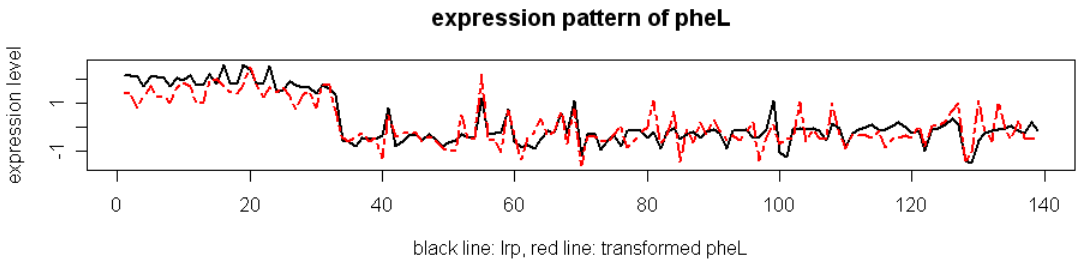


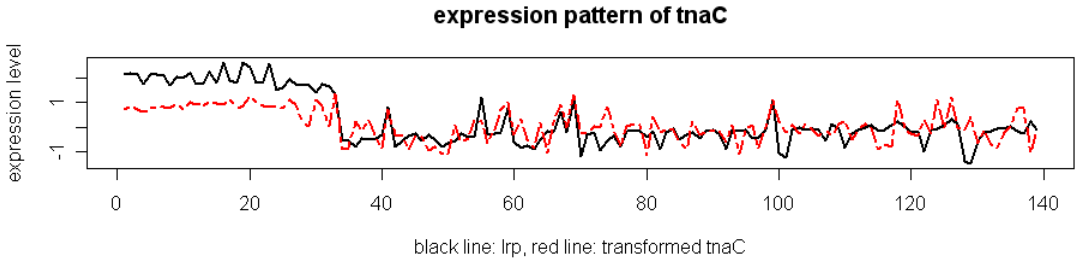

expression pattern of pyrB
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed pyrB

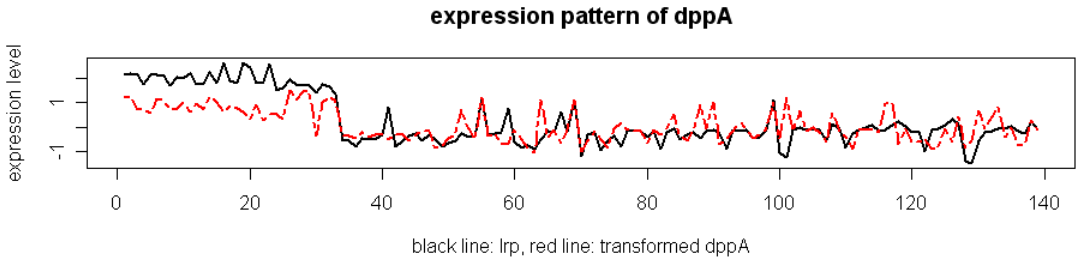

expression pattern of yagU
expression level
-1
1
3
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed yagU

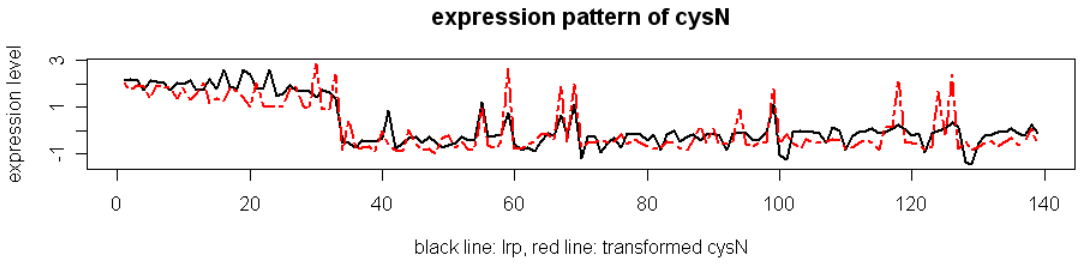

expression pattern of cysD
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed cysD

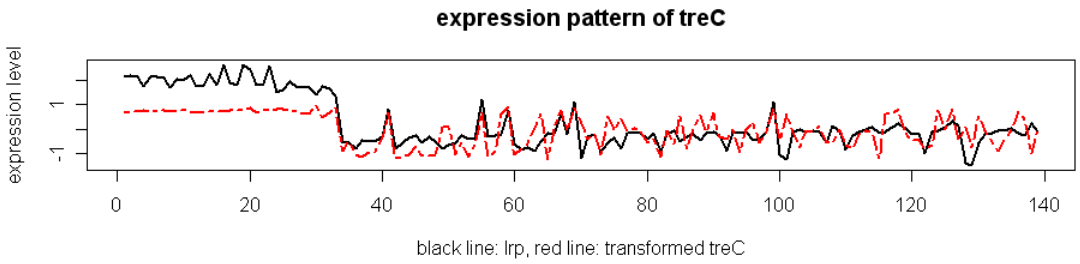

expression pattern of cyK
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed cyK

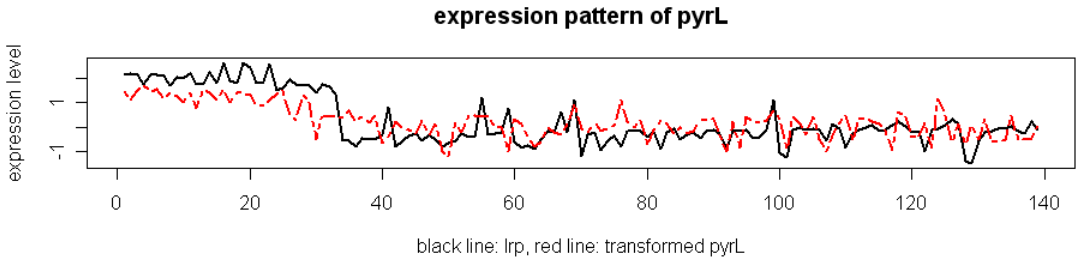

expression pattern of pheL
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed pheL

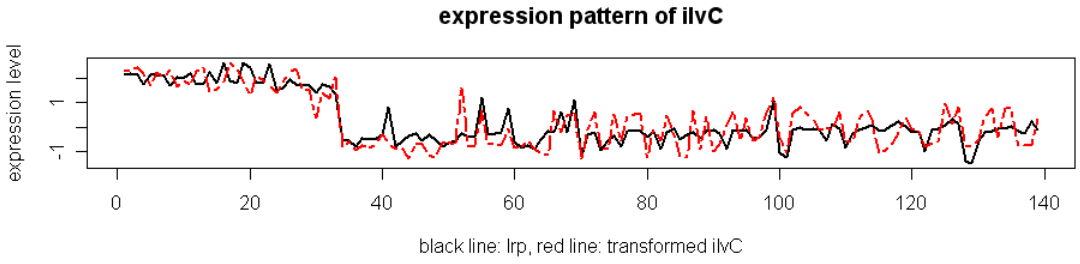

expression pattern of tnaC
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed tnaC

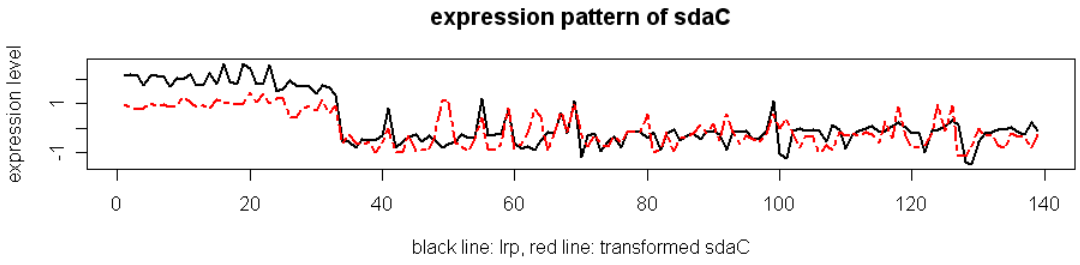

expression pattern of dppA
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed dppA

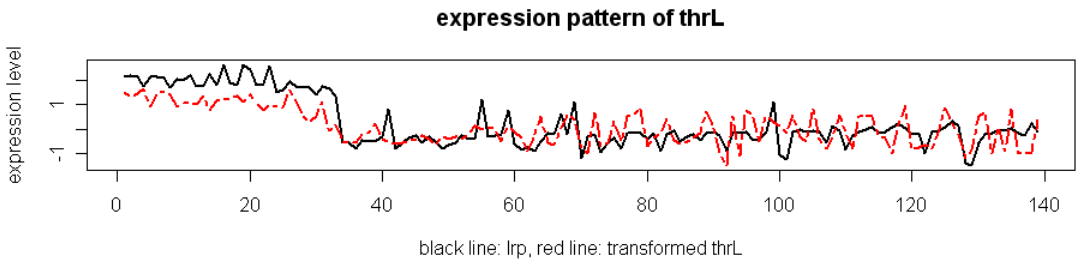

expression pattern of cysN
expression level
-1
1
3
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed cysN

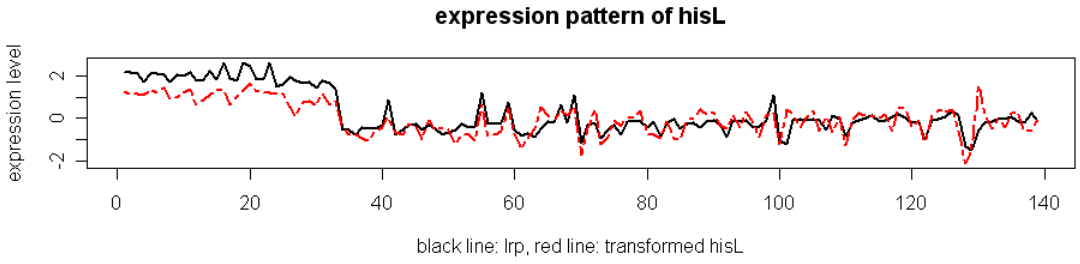

expression pattern of treC
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed treC

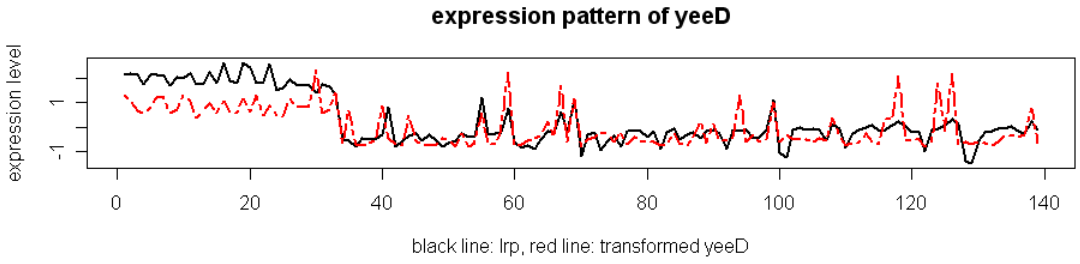

expression pattern of pyrL
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed pyrL

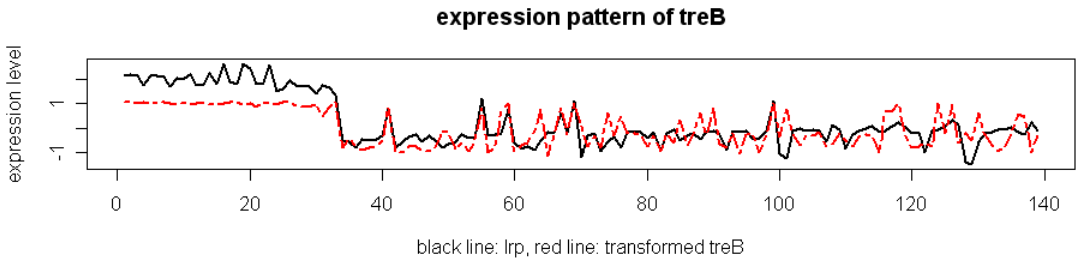

expression pattern of ilvC
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed ilvC

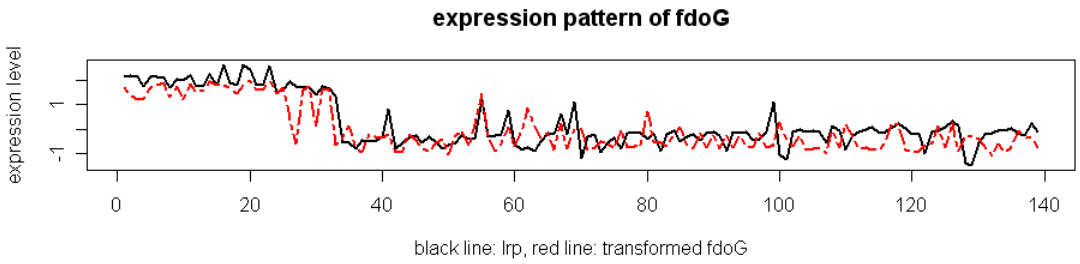

expression pattern of sdaC
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed sdaC

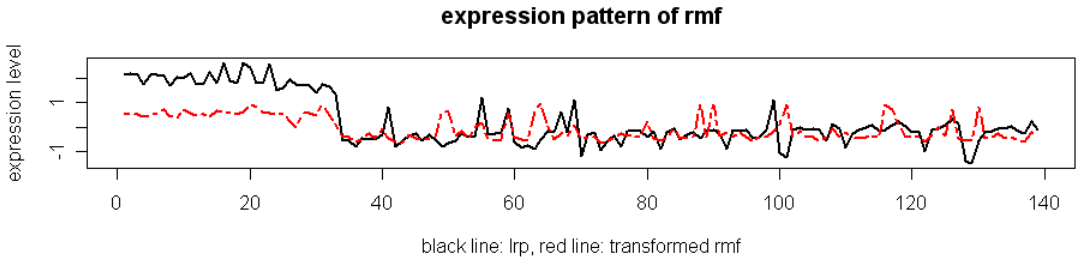

expression pattern of thrL
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed thrL

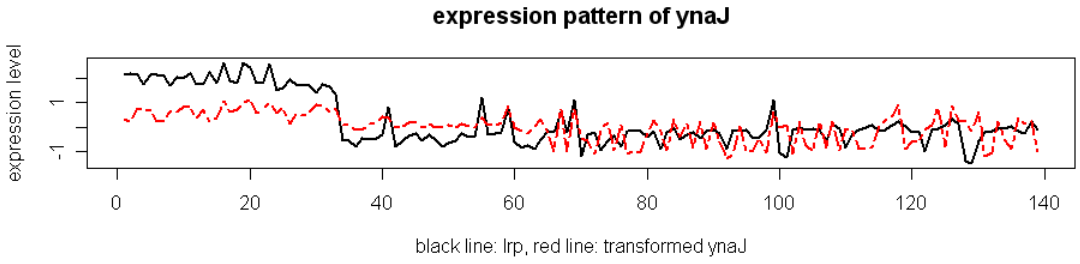

expression pattern of hisL
expression level
-2
0
2
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed hisL


expression pattern of yeeD
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed yeeD


expression pattern of treB
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed treB


expression pattern of fdoG
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed fdoG


expression pattern of rmf
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed rmf


expression pattern of ynaJ
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed ynaJ

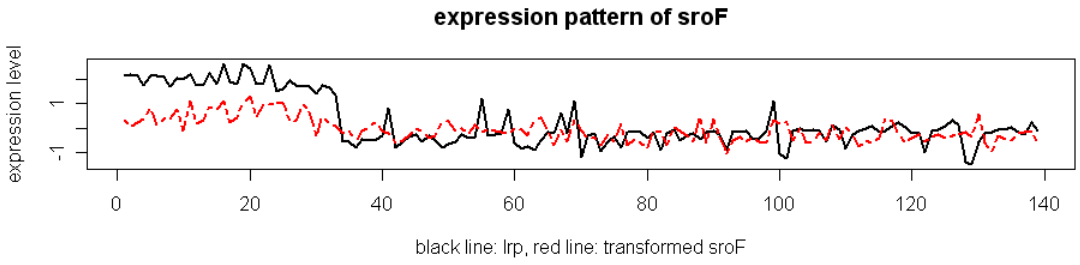
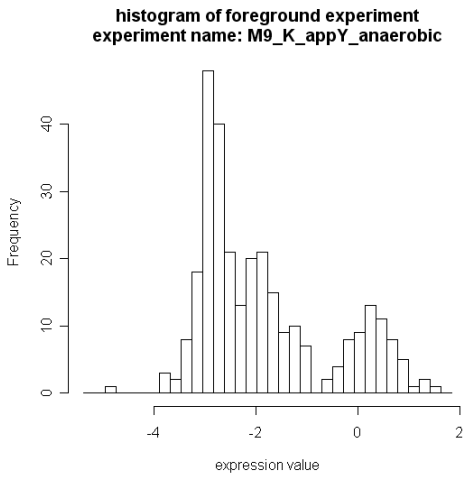
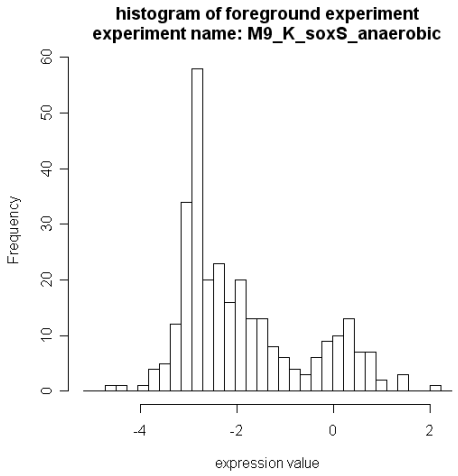

expression pattern of sroF
expression level
-1
1
0
20
40
60
80
100
120
140
black line: lrp, red line: transformed sroF

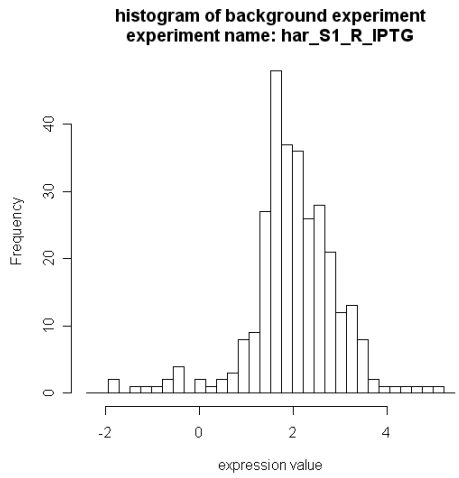
Figure A2.9 A. Histograms of expression profile differences (z_{ij}) in the top five experimental conditions (foreground). B. Histogram of expression profile differences (z_{ij}) in the bottom five experimental conditions (background). Data used here is the 300-gene test set selected from the *E. coli* microarray compendium.

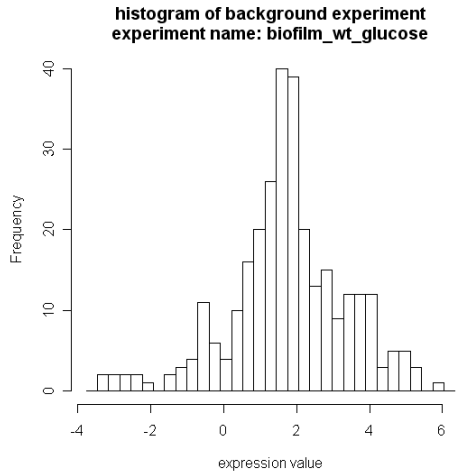
A.





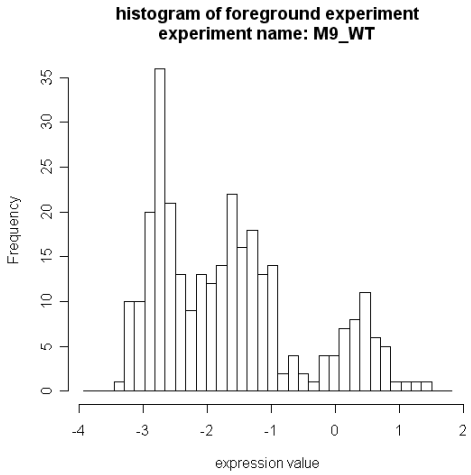
B.

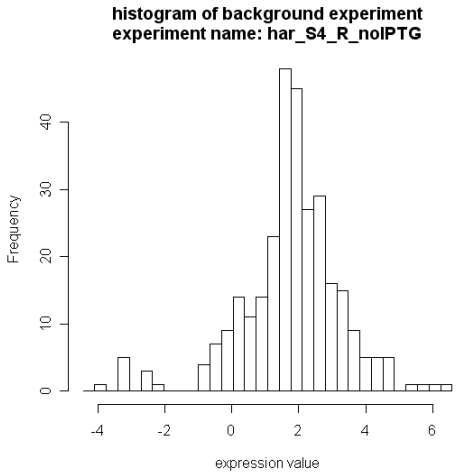


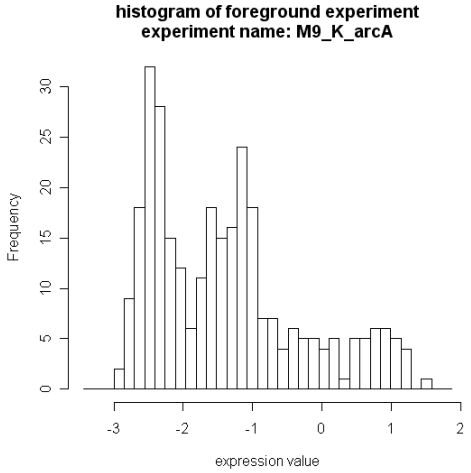


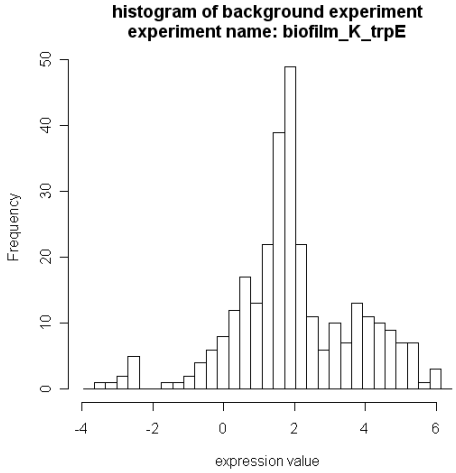
A.

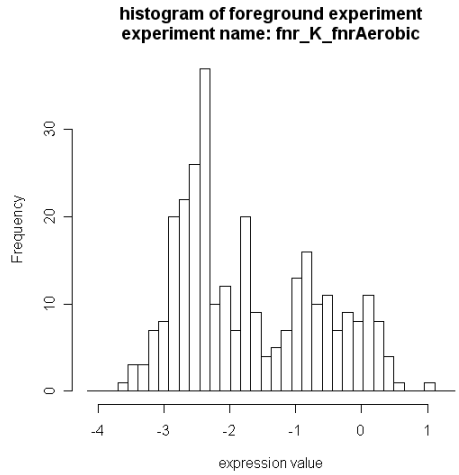
B.

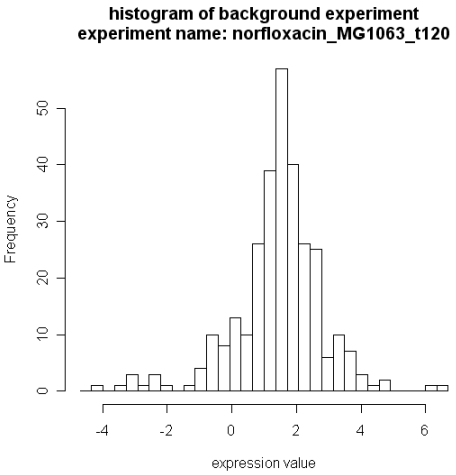

histogram of foreground experiment
experiment name: M9_WT
Frequency
expression value


histogram of background experiment
experiment name: har_S4_R_nolPTG
Frequency
expression value


histogram of foreground experiment
experiment name: M9_K_arcA
Frequency
expression value


histogram of background experiment
experiment name: biofilm_K_trpE
Frequency
expression value


histogram of foreground experiment
experiment name: fnr_K_fnrAerobic
Frequency
expression value


histogram of background experiment
experiment name: norfloxacin_MG1063_t120
Frequency
expression value

CHAPTER 3

On the detection and refinement of transcription factor binding sites using ChIP-Seq data

3.1 Introduction

Accurately locating the transcription factor (TF)-DNA interaction sites provides key insights into the underlying mechanisms of transcriptional regulation. By exploiting the fact that binding sites for a specific TF often show sequence specificity, computational prediction of TF binding sites, or motif finding, has become an indispensible tool for functional genomics research. A variety of different software programs have been developed for motif-finding (44,54-59) (see Tompa et al. (60) for a review of this topic).

The input data for computational motif-finding algorithms are DNA sequences believed to be enriched by the TF binding sites, or motifs. Typical sources of the input data are known co-regulated genes (59), phylogenetic conservation (61), or results from functional genomics experimental assays (54,62-64). For the latter, continually evolving high-throughput technologies, from DNA microarray (1,11) to ChIP-chip (65,66) and now ChIP-Seq (4,67-69), offer rapidly improving opportunities for motif finding.

ChIP-Seq, or chromatin immunoprecipitation (ChIP) (70,71) followed by ultra-high-throughput sequencing, has emerged as a powerful new technology for genome-wide mapping of protein-DNA interactions and histone modifications (4,67-69). Through direct sequencing of all DNA fragments from ChIP assays, ChIP-Seq can reveal protein-DNA interaction sites across the entire genome, thus building a comprehensive and high-resolution interactome map for DNA-binding proteins of interest.

From past experience, exploiting the quantitative information provided by high-throughput genomic assays allows scientists to develop more effective motif-finding algorithms. Improvements in motif detection have been reported in studies using microarray (62,63) and ChIP-chip (54,64) data. The newly emerged ChIP-Seq technology has demonstrated remarkable sensitivity and specificity in identifying protein-DNA binding loci across the entire genome with high resolution and few constraints. In excess of 10,000 DNA sequences are routinely being identified as candidates that potentially harbor protein-DNA interaction sites of interest. Such information provides an exciting new venue for motif discovery and refinement.

A *de novo* motif search is a natural follow-up to the identification of ChIP-enriched regions. Not only is it required when the TF binding motif pattern is unknown, it is also important in cases where TF and its canonical binding motif pattern have been established. After all, it is reassuring to be able to "rediscover" the known TFBS motif pattern from the input sequences. More importantly, most of the known TF binding motif patterns stored in the various TF binding motif databases or reported in the literature are defined based on limited numbers of experimentally verified TF-DNA interaction sites. Many of these motif patterns could be inaccurate due to limited experimental data.

Performing a *de novo* motif search on a large number of ChIP-Seq binding sites has the potential to refine the motif patterns of the TFBS.

While a variety of methods that attempt to identify ChIP-enriched genomic regions from ChIP-Seq experiments (also called "peak calling") have been described (72-80), little has been developed utilizing ChIP-Seq data for motif finding.

Probability model-based *de novo* motif finding algorithms such as MEME have demonstrated a high level of sensitivity and specificity (44,55-57,81-85). However, since these methods were developed when only a handful of motif-enriched sequences were available, they do not work well when analyzing large sets of sequences identified by ChIP-Seq. There are at least two limitations that affect their performance: (1) the requirement for going through all bases in all sequences using time-consuming iterative procedures means that these methods do not scale well for the analysis of large sets of sequences generated from ChIP-Seq; (2) existing methods, which only consider sequence data, are unable to fully utilize the rich information produced from ChIP-Seq. Overlooked information includes the sequencing depth along the ChIP-enriched regions and the overall significance of ChIP-enrichment for each sequence. "Sequencing depth" refers to the number of ChIP DNA fragments that cover each base. Currently, a common practice for performing motif finding on ChIP-Seq data is to use existing motif-finding tools on a subset of all sequences (e.g., the top 500 sequences or top 10% of all such sequences) (74,75). This is sub-optimal because the small sample size may lead to an inaccurate motif pattern and the selection of top sequences tends to result in motif patterns that are overly-conserved.

We believe that a more desirable approach is to develop algorithms that can utilize all of the sequence information generated from ChIP-Seq. Not only will this strategy result in the identification of more accurate motif patterns, but we also predict that the dramatically increased number of *in vivo* binding sites revealed by ChIP-Seq will permit the use of probability models that are more sophisticated than the commonly-used product multinomial models (83) for characterizing the motif pattern.

To address these limitations and fully exploit the information provided by ChIP-Seq experiments, we develop a novel model-based motif-finding algorithm named the Hybrid Motif Sampler (HMS). It is specifically designed for ChIP-Seq data and utilizes *all* ChIP-enriched regions identified from ChIP-Seq experiments. In this algorithm, we propose a new probability model that considers both DNA sequence and sequencing depth information that is available from ChIP-Seq experiment. It also allows inter-dependent positions within a motif to be identified. In addition, we propose a novel hybrid searching scheme to significantly expedite the iterative procedure. Our algorithm is capable of processing tens of thousands of sequences and is much faster than the established *de novo* motif-finding tools such as MEME.

3.2 Methods

3.2.1 The statistical model

Let $R = (R_1, \ldots, R_J)$ denote a set of J sequences (e.g., DNA sequences in ChIP-enriched regions identified by ChIP-Seq) of length $L_1, \ldots, L_J$. We initially assume that every sequence R_j contains exactly one binding site. In addition, the vector that is formed by the start locations is referred to as the alignment variable, denoted as $A = (A_1, \ldots, A_J)$ where $1 \leq a_j \leq L_j - w + 1, j = 1,2, \ldots J$. Here w is the motif width and is assumed to be

known. Given A and w, the aligned sequence motif can be represented by a four by w matrix. Each column of the matrix stores the frequency counts of the four types of nucleotides. Liu et al. (83) proposed the *product-multinomial model* to model the nucleotide preferences shown in such matrices. The product-multinomial model has been widely used in EM-based (57,81) and Gibbs sampler-based (56,82,84) motif finding algorithms. Let $\Theta = (\theta_1, \ldots, \theta_w)$, θ_i represent the nucleotide preference at the i th position of the motif and let the probability vector θ_0 represent the nucleotide preference for non-motif positions in these sequences. Each of the $\theta_i, i = 0,1,2,\ldots,w$ is a probability vector of length four. For notational simplicity, we use integers 1, 2, 3 and 4 to represent the four types of nucleotides A, C, G and T.

For *de novo* motif finding, the parameter of main interest in our model is the alignment variable A. Lawrence et al. (56) proposed a Gibbs sampler-based approach in which the posterior distribution for alignment a_j can be expressed as:

$$p(a_j = l \mid \theta_0, \Theta, R_j, A_{-j}) \propto \prod_{k=1}^{4} \theta_{0k}^{h_k(R_j)} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})}$$

$$\propto \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})} \quad (3.1)$$

Where $A_j = (a_1, \ldots a_{j-1}, a_{j+1}, \ldots a_J)$ and the functions $h_k(), k = 1,2,3,4$, returns the number of nucleotides of type k.

For θ_0 and Θ, an alternative to sampling them from posterior conditional distributions as in a standard Gibbs sampler, one can use the predictive updating technique (83) to integrate them out. Alternatively, the posterior means can be used to

approximate the updated parameters during iteration. More details of these strategies can be found in Liu et al. (83)

3.2.2 Allowance for some sequences that do not contain the motif

In the model above, we assume that every sequence R_j contains exactly one motif. However, this is not the case in real ChIP-Seq data. To increase specificity, it is highly desirable that we generalize the method to allow some sequences to be motif-free. We introduce a binary indicator variable I_j where $I_j = 1$ indicates that R_j contains at least one motif, and $I_j = 0$ otherwise. In the algorithm, I_j is set to 1 if the average of likelihood ratios observing the motif in the sequence R_j, denoted as z_j, is greater than 1. i.e.,

$$z_j = \frac{1}{L_j - w + 1} \sum_{l=1}^{L_j - w + 1} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1})}, I_j = I_{\{z_j \geq 1\}} \tag{3.2}$$

After updating I_j, we only conduct motif search on the sequences with $I_j = 1$.

3.2.3 Modeling sequencing depth

The model described in equation (3.1) assumes that binding motifs are equally likely to occur at all positions in each sequence. This is reasonable when no information beyond the input DNA sequences is considered. However, such a model is no longer sufficient for analyzing ChIP-Seq data since additional information beyond the DNA sequences is available and should be incorporated. In particular, it has been shown that the sequencing depth in each ChIP-enriched region is indicative of the motif location (74,77). Figure A3.1 in appendix shows that: the majority of motifs are tightly packed near the peak summit (the location inside each peak with the highest sequence coverage depth), especially for the highly-significant peaks.

To capitalize on the extra information provided by ChIP-Seq, we propose adding to the method an informative prior distribution of the motif location based on the sequencing depth. There are multiple ways to assign such priors. The simplest strategy is to make the prior probabilities directly proportional to the sequencing depth in each sequence. However, since sequencing depth is affected by many factors, such as local GC content, using a prior distribution like this may result in "over fitting". Alternatively, a parametric distribution that approximates the sequencing depth can be used to obtain the prior probabilities. In this study, we set the prior probabilities to be proportional to a discretized Student's t-distribution with three degrees of freedom and rescaled such that the prior probabilities form a step function with a fixed step-size (25 bp in this study). The prior probabilities are symmetric and centered at the peak summit (most peak-calling software provides the exact location of the summit). Specifically, the prior probabilities that a motif starts at position l can be expressed as:

$$p(a_j = l) \propto t_3\left(int\left[\frac{|l + w/2 - s_j| + u/2}{u}\right]\right) \tag{3.3}$$

Where t_3 is the probability density function of the Student's t-distribution with three degrees of freedom, s_j is the location of the peak summit, w is the motif width, u is the step size (25 bp in this study) in the step function and int [·] returns the integer part of a real number. Please see Figure A3.2 in appendix for an illustration of the prior probabilities. The reason that we choose Student's t-distribution instead of a normal distribution is because it better allows for some motif locations to be far from the peak (the standard deviation of Student's t-distribution with three degrees of freedom is 1.73, compared to one for standard normal distribution).

3.2.4 Modeling intra-motif dependency

The classical product-multinomial model assumes that the positions within the motif are independent of each other (86). However, recent studies indicate that some positions of TF binding motifs exert an inter-dependent effect on the binding affinities of TF's (87-90). These findings imply that the commonly used product-multinomial model may be too simplistic in characterizing the binding sites. Models that allow for dependent positions likely will provide a better fit of the data. The significantly increased quantity of motifs identified by ChIP-Seq enables us to consider a more sophisticated model that can take into account the intra-motif dependency.

There have been numerous attempts to incorporate into models the inter-dependency among positions within a motif. King and Roth (91) introduced a non-parametric representation of motifs that allows arbitrary dependencies among positions. Barash et al. (92) suggested multiple Bayesian network models to represent dependencies among motif positions. Zhou and Liu (93) proposed a generalized weight matrix model in which a 16-component multinomial model is used to model two dependent positions jointly.

Here we extend the generalized weight matrix model of Zhou and Liu. To take greater advantage of the abundant sequence information made available by the ChIP-Seq technology, our model allows up to three positions to be inter-dependent.

3.2.5 Detection of dependent positions

Given a set of aligned putative binding motifs, our goal is to identify positions that show inter-dependency. Here "inter-dependency" implies conservation of nucleotide combinations spanning multiple positions that cannot be accurately described by the

independent motif model. As an example, for a pair of positions, if the frequency of a particular dinucleotide, say AC, is much higher or lower than the product of frequency of nucleotide A in the first position and frequency of nucleotide C in the second position, we conclude that the two positions are dependent.

A variety of methods have been proposed in the literature to search for such inter-dependent positions. Barash et al. (92) applied machine learning approaches to infer the structure of a Bayesian network that best represents the underlying motif. Zhou and Liu (93) proposed a Metropolis-type iterative procedure to identify pairs of inter-dependent positions. Given the abundant motif data from ChIP-Seq, we implement a comprehensive search strategy to go through all pairs of positions within the motif to determine whether there is evidence of dependency. To be specific, for any two positions i and j among $w(w-1)/2$ possible pairs, we first obtain probability estimates of the 16 dinucleotides assuming either a 16-component multinomial model (dependent) or the product of two four-component multinomial models (independent). Let the number of motifs be represented by M. The term $g_x(r_i)$ represents the number of motifs whose i th position is occupied by nucleotide x and the term $g_{xy}(r_i, r_j)$ represents the number of motifs whose i th and j th positions are occupied by nucleotides x and y respectively. The probability estimates under the two competing models are $\hat{\eta}_x(r_i) = g_x(r_i)/M$ and $\hat{\eta}_{xy}(r_i, r_j) = g_{xy}(r_i, r_j)/M$, respectively. We then calculate the Hamming distance between the two sets of estimates as

$$d_{ij} = \sum_{x=1}^{4} \sum_{y=1}^{4} |\hat{\eta}_{xy}(r_i, r_j) - \hat{\eta}_x(r_i)\hat{\eta}_y(r_j)| \tag{3.4}$$

Under the hypothesis that the two positions are independent, we expect that distance $d_{ij} = 0$, excluding sampling variability; larger d_{ij} indicates stronger inter-dependency between positions i and j. In this study, we designate positions i and j to be dependent if $d_{ij} > 0.2$. The threshold is determined from the empirical null distribution of d_{ij} inferred through simulations. More details can be found in appendix.

3.2.6 Posterior distribution

We take a Bayesian approach and consider two different models to describe the motif pattern. In the first one, we assume all positions within the motif are independent. There are two sets of parameters in this model: alignment variable A and multinomial distribution probability vector $\theta_i, i = 0,1,\dots,w$. The prior distributions for A are multinomial with probabilities defined as in equation (3.3). Adopting a conjugate prior distribution for each θ_i, which is $Dirichlet(\alpha_{0,1},\dots,\alpha_{0,4})$, the posterior probabilities that a motif starts at position l can be expressed as:

$$p(a_j = l \mid \theta_0, \Theta, R_j, A_{-j}) \propto I_{\{z_j > 1\}} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}}\right)^{h_k(r_{j,l+i-1}) + \alpha_{0,k}} p(a_j = l) \qquad (3.5)$$

As suggested in Liu et al. (83), the above conditional distribution can be closely approximated by replacing θ_{ik} by its posterior mean given the current alignment vector A_{-j}:

$$\theta_{ik} = \frac{h_k(r_{j,l+i-1}) + \alpha_{0,k}}{N_{0[-j]} + \alpha_{0,k}}, N_{0[-j]} = \sum_{s \neq j} I_{\{z_j > 1\}} \qquad (3.6)$$

For background (non-motif) regions, it has been shown that employing a Markov model to capture weak dependency in background DNA sequences improves the sensitivity and specificity of motif finding compared to an independent model in equation (3.1). In this

study, we use a third-order Markov model as in Liu et al. (55) to characterize the background sequences. Under such a model, the probability of observing DNA sequence fragment $\{r_{s,t}, r_{s,t+1}, \cdots, r_{s,t+w-1}\}$ in the background can be represented by

$$P(Background_{s,t}) = P(r_{s,t})P(r_{s,t+1}|r_{s,t})P(r_{s,t+2}|r_{s,t+1}, r_{s,t})P(r_{s,t+3}|r_{s,t+2}, r_{s,t+1}, r_{s,t}) P(r_{s,t+4}|r_{s,t+3}, r_{s,t+2}, r_{s,t+1}) \cdots \quad (3.7)$$

In this background model, the $3 \times 4^3 = 192$ conditional probabilities are estimated from human promoter sequences downloaded from UCSC genome browser website (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). The dataset contains 5kb upstream sequences of annotated transcription starts for all RefSeq genes with annotated 5' UTRs.

After incorporating these modifications, the complete posterior distribution for $a_j = l$ becomes

$$p(a_j = l|\theta_0, \Theta, R_j, A_{-j}) \propto I_{\{z_j>1\}} \frac{\prod_{i=1}^{w}\prod_{k=1}^{4} \hat{\theta}_{ik}^{\,h_k(r_{j,l+i-1})+\alpha_{0,k}}}{P(Background_{s,t})} p(a_j = l) \quad (3.8)$$

In the second model, we consider intra-motif dependency. Within the motif, we assign positions into two disjoint groups: groups of independent positions S and groups of dependent position pairs P where $P = \{(i,j): d_{ij} > 0.2\}$. By modeling dependent positions jointly, the probability "matrix" Θ becomes an amalgam of vectors of length four (modeling single positions) and vectors of length 16 (modeling pairs of dependent positions). The prior distributions for the two types of θ_j's are $Dirichlet(\alpha_{0,1}, \cdots, \alpha_{0,4})$ and $Dirichlet(\beta_{0,1,1}, \cdots, \beta_{0,1,4}, \beta_{0,2,1}, \cdots, \beta_{0,4,4})$ respectively. The complete posterior distribution for $a_j = l$ in the dependent model is

$$p(a_j = l|\theta_0, \Theta, R_j, A_{-j}) \propto I_{\{z_j>1\}} \frac{U \cdot V}{P(Background_{s,t})} p(a_j = l)$$

$$U = \prod_{i \in S} \prod_{k=1}^{4} \hat{\theta}_{ik}^{\,h_k(r_{j,l+i-1})+\alpha_{0,k}}$$

$$V = \prod_{i_1,i_2 \in P} \prod_{k_1=1}^{4} \prod_{k_2=1}^{4} \hat{\theta}_{i_1 i_2}^{\,h_{k_1 k_2}(r_{j,l+i_1-1}, r_{j,l+i_2-1})+\beta_{0,k_1,k_2}} \quad (3.9)$$

Here the counting function $h_{k_1 k_2}()$, whose argument is a set of positions, counts the frequency of the 16 dinucleotides for a pair of positions within the motif. The above model can be extended easily to allow three-way inter-dependent positions.

3.2.7 Acceleration via prioritized hybrid Monte Carlo

To streamline this motif-finding algorithm in order to handle a large number of input sequences, we develop a prioritized hybrid strategy to increase computation speed with only minimal if any sacrifice in accuracy. Unlike a standard Gibbs sampler where motif alignment variables are sampled stochastically from *all* sequences, only a small proportion, π, of all sequences are subjected to stochastic sampling. For the remaining sequences, we select the alignment variable deterministically by identifying the position that corresponds to the highest probability as given by equation (3.8) or (3.9). Since the deterministic approach is much faster than the stochastic one and the proportion π we use is often quite small (π ~10%), this hybrid strategy is much faster than the standard Gibbs motif sampler (56).

For each iteration, the proportion of sequences undergoing stochastic search is constant, but a different set of sequences is selected each time. We have automated the process of selecting a subset of sequences for stochastic search. All the sequences identified from the ChIP-Seq experiment are rank-ordered according to their ChIP-enrichment. Assume we run N iterations in each Gibbs sampler. In the i th iteration, we

sample a fixed number of $\pi \times J$ sequences from a multinomial distribution $mult(J, p_{i1}, \ldots p_{iJ})$. At the beginning of the iteration, we use a monotonically decreasing triangle probability distribution, which assigns higher probability to sequences with higher ChIP-enrichment. As the iteration proceeds, the slope of the triangle gradually becomes flatter so that the oversampling of higher ChIP-enriched sequences diminishes. In the last iteration, the distribution becomes uniform. For the i th iteration, we have

$$p_{ij} \propto c_{ij} = J - j + 1 - \frac{J/2 - j + 1}{N - 1} \times (i - 1), i = 1, \ldots, N; j = 1, \ldots, J. \quad (3.10)$$

3.2.8 Implementation

We have developed a software program that implements the algorithms described in this manuscript. The HMS program is a Gibbs sampler type iterative procedure. To reduce the possibility that the Markov chain converged to a local mode, we run multiple Markov chains and choose the motif pattern that corresponds to the highest likelihood as the final motif pattern. The number of parallel chains and the number of complete iterative cycles within each chain are specified by users. Within each chain, the iterative procedure can be broken down into three steps. In the first step, we use a traditional product multinomial model in which all positions are assumed independent of each other. We further assume every sequence contains one motif. In the second step, we again assume all positions are independent, but we allow some sequences to be motif-free. In the final step, we adopt the generalized motif model that allows intra-motif dependency. The HMS program, including the source code is freely available at http://www.sph.umich.edu/csg/qin/HMS/.

3.2.9 Performance evaluation using simulated data

In the simulation study, we are interested in evaluating the performance of various *de novo* motif finding algorithms from two perspectives: first, the number of times a program successfully detects the motif inserted into each of the 100 simulated datasets; second, the accuracy of the inferred motif pattern given that the motif has been found.

For the former, since we know the true location of all inserted motifs in the simulated datasets, we are able to directly verify whether each motif site predicted by the testing software is correct. Within each simulated dataset, we declare that the inserted motif is found if the proportion of sequences in which the program correctly identifies the true motif location is greater than 20%.

For the latter, we measure the accuracy of an inferred motif pattern by calculating the average Hamming distance between the true probability matrix Θ and its prediction denoted as $\hat{\Theta}$:

$$h = \frac{1}{w}\sum_{i=1}^{4}\sum_{j=1}^{w}|\theta_{ij} - \hat{\theta}_{ij}| \tag{3.11}$$

3.2.10 Performance evaluation using real data

Given a set of sequences identified by ChIP-Seq, we want to discern which *de novo* motif-finding algorithm produces a more accurate motif pattern. Since the exact true motif pattern is unknown, we use motif enrichment as the criterion. We assume that among multiple motif patterns, the one that is most enriched in the ChIP-Seq-identified regions relative to random controls is closest to the true motif pattern.

We use a cross-validation scheme to assess motif enrichment. The original dataset is equally divided into halves: a training set and a testing set. The input sequences are

restricted to within 200 bp in length and centered at the peak summit (no more than 100 bp toward each side of the peak summit). For the testing set, we create a control set composed of randomly selected DNA promoter sequences (within 5kb upstream of the transcription start site) as in Zhou and Liu (93) matched by number of sequences and length of each sequence. We run each motif-finding program on the training set to identify the motif pattern, and then utilize this pattern to scan both the testing and the corresponding control sets to assess how many sequences contain the motif. We employ a set of significance thresholds and calculate the corresponding empirical false discovery rate (FDR) (94) and motif enrichment, as measured by Chi-square test statistics for a 2×2 contingency table. The empirical FDR is estimated by dividing the number of control sequences that contain the motif by the number of testing sequences that contain the motif. We repeat the scheme five times for each dataset and report the average test statistics corresponding to each FDR level.

We plot the curves of the empirical FDR versus the Chi-square test statistics when the empirical FDR is between 0 and 0.2. To accomplish this, we equally divide the empirical FDR into ten consecutive windows and calculate the mean of the Chi-square test statistics from five cross validations (when the corresponding empirical FDRs fall into the same window). Since the curve representing the most enriched motif pattern will be the highest, we use area under the curve (AUC) as a quantitative assessment of the overall motif enrichment. Higher AUC indicates further motif enrichment.

3.2.11 Estrogen receptor ChIP-Seq experiment on MCF7 cells

To test the algorithms in a real setting, we have conducted a ChIP-Seq experiment to survey genome-wide binding of Estrogen Receptor (ER) on the MCF-7 breast cancer

cell line. ER is a hormonal TF that, when liganded by estrogen, binds specially to estrogen response elements (ERE) and plays a critical role in breast cancer development. Identifying ER target genes and refining the ERE motifs are thus of significant interest. A brief description of the experimental protocol is shown in the next paragraph. More details can be found in appendix.

Briefly, MCF-7 cells were grown in RPMI media supplemented with 10% FBS to 50% confluence. The cells were then hormone-starved for three days prior to treatment of the vehicle control or 10nM β–estradiol for 45 minutes. The cells were then harvested for ChIP analysis using an antibody against estrogen receptor (ER)-alpha (sc-543x, Santa Cruz) or against IgG. The ChIP-enriched DNA was evaluated for significant enrichment of positive control genes and then subjected to ChIP-Seq sample preparation and short-read sequencing using Illumina Genome Analyzer (Illumina Inc, San Diego, CA) following the manufacturer's protocols. The raw sequencing images were analyzed using the Illumina analysis pipeline, and the sequencing reads were subsequently aligned to the human reference genome (NCBI v36, hg18) using ELAND software (Illumina Inc, San Diego, CA), producing sequencing reads of 35 bps. Only sequencing reads that are uniquely mapped to the human reference genome with up to two mismatches were included for further analysis as delineated in this study. We used the HPeak software program (95), to define the ChIP-enriched regions. Details of the HPeak software program can be found in appendix.

3.3 Results

3.3.1 Simulation study

3.3.1.1 Independent motif models

The goal of this simulation study was to evaluate the ability of HMS to identify the correct motif patterns. We use the default setting for HMS which adopts the informative prior and allows intra-motif dependency. For comparison, we also tested a simpler version of HMS that assumes all positions are independent. In addition, we applied two established motif-finding software tools, MDscan (54) and MEME (57) on the same sets of simulated data. Following the simulation scheme of Liu et al. (54), four motif models were manually created (Table A3.1A), representing two different motif widths (8 bps and 16 bps), and two different degrees of conservation measured by information content (1.42 and 0.93). The information content is defined as:

$$\frac{1}{w}\sum_{i=1}^{4}\sum_{j=1}^{w} p_{ij}\log_2(4p_{ij}) \tag{3.12}$$

where p_{ij} is the proportion of base j at the motif position i. Information content ranges from 0 to 2, reflecting the weakest to the strongest motifs. Finally, two different motif abundance schemes (Table A3.1B) were considered for a total of eight combinations in the simulation study. The eight simulation settings covered a wide range of scenarios. The combination of short motif width, weak motif information content and low motif abundance was the most challenging.

For each setting, we simulated 100 test datasets. Each dataset contains 3,000 sequences of 200 bp in length. To mimic real human data, all the sequences were generated from a third-order Markov model with parameters estimated from the

collection of 5kb promoter sequences of annotated genes in the human genome (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). Hypothetical motifs were generated from product multinomial models with specified length and information content. The proportion of sequences that contained a motif followed one of the two abundance schemes mentioned in the previous paragraph. We assumed that each sequence contained at most one motif.

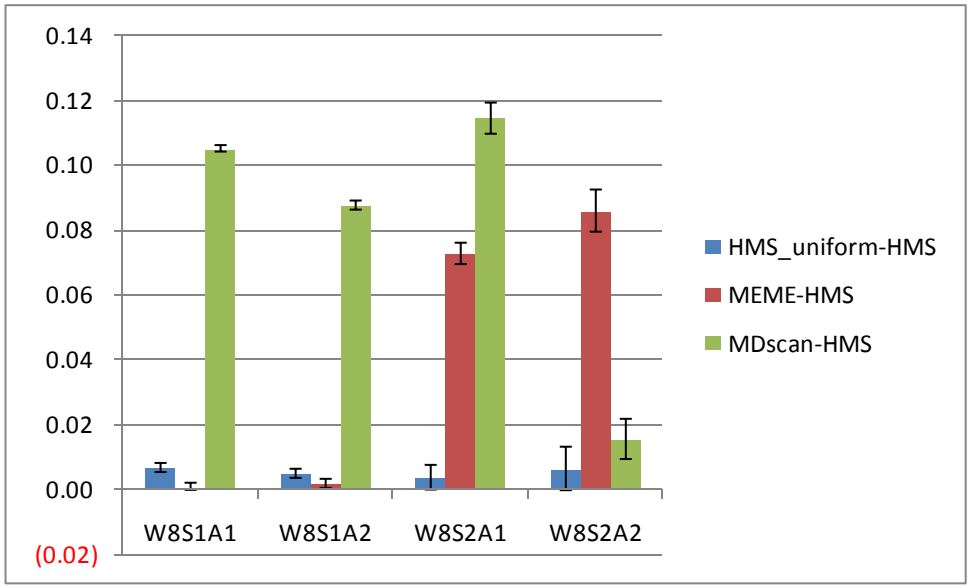
We next derived the empirical distribution from real ChIP-Seq data of CTCF and NRSF of the motif start locations in a 200 bp window centered at the peak summit. We strategically inserted the motifs in these sequences following this empirical distribution. As a consequence, the motif locations were biased toward the center of the sequence, which was assumed to be the location of the peak summit.

We applied MDscan, MEME and HMS to every dataset. Two versions of HMS were used in the comparison. One assumed an informative prior (proportional to a discretized and rescaled Student's t-distribution with three degrees of freedom) that favored motif start locations near the peak. The other, denoted as HMS_uniform, assumed a uniform prior for the motif start location throughout the genome. As described in the Methods section, we used the successful motif detection rate and the accuracy of predicted motif pattern as measurements of performance.

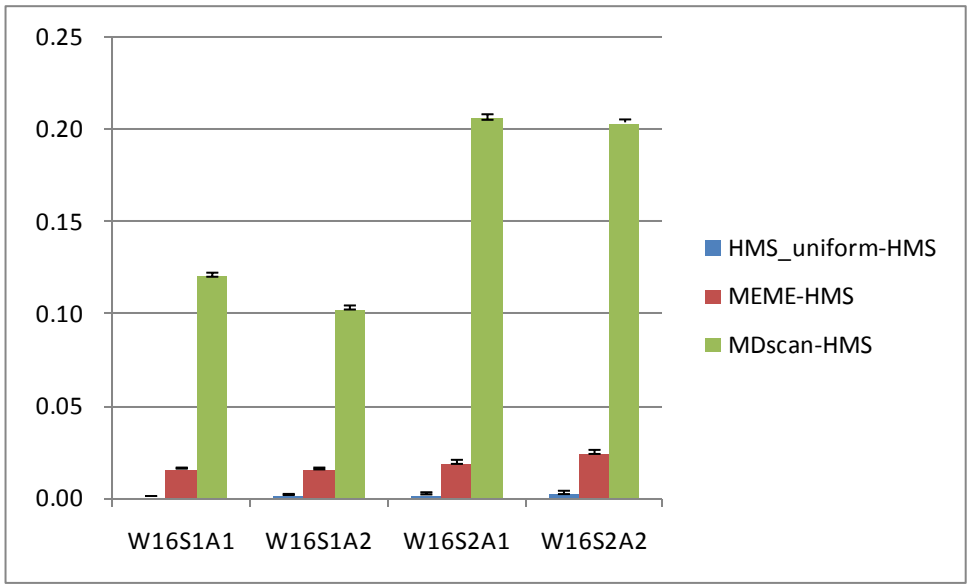
For the motif detection rate, both versions of HMS achieved perfect results in all eight simulation settings. MEME and MDscan achieved perfect results in six and four settings respectively. MEME achieves equal or higher detection rate than MDscan in all but one setting. (Table A3.2A in appendix).

Figure 3.1 Performance comparison on simulated data with independent and dependent motif model. The y-axis represents the difference between two sets of discrepancies resulted from two different motif finding methods. The discrepancies are measured by the average Hamming distance between the estimated and the true probability matrix Θ. A. Independent, motif width = 8 bp. B. Independent, motif width = 16 bp. C. Dependent, motif width = 8 bp. D. Dependent, motif width = 16 bp.

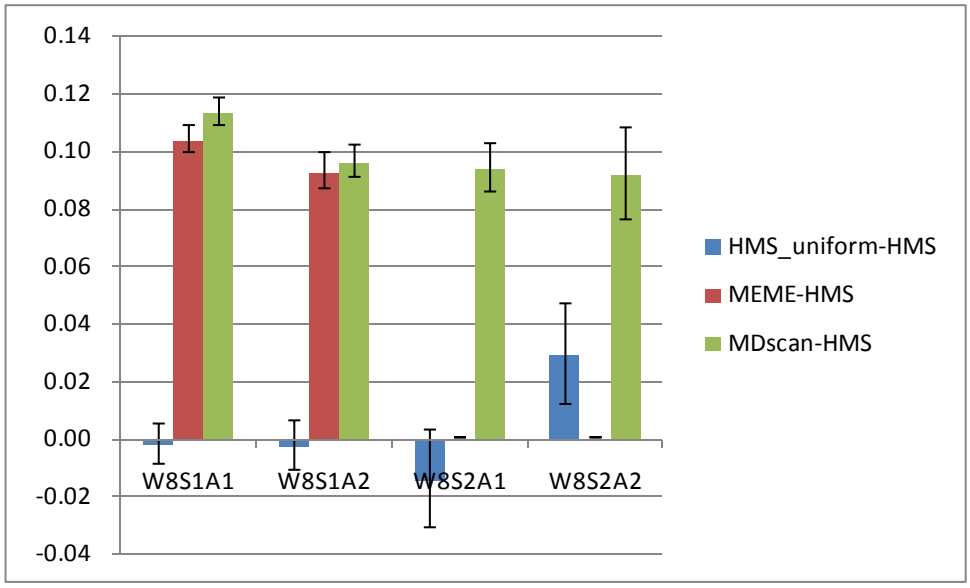
A. Independent, motif width = 8 bp.



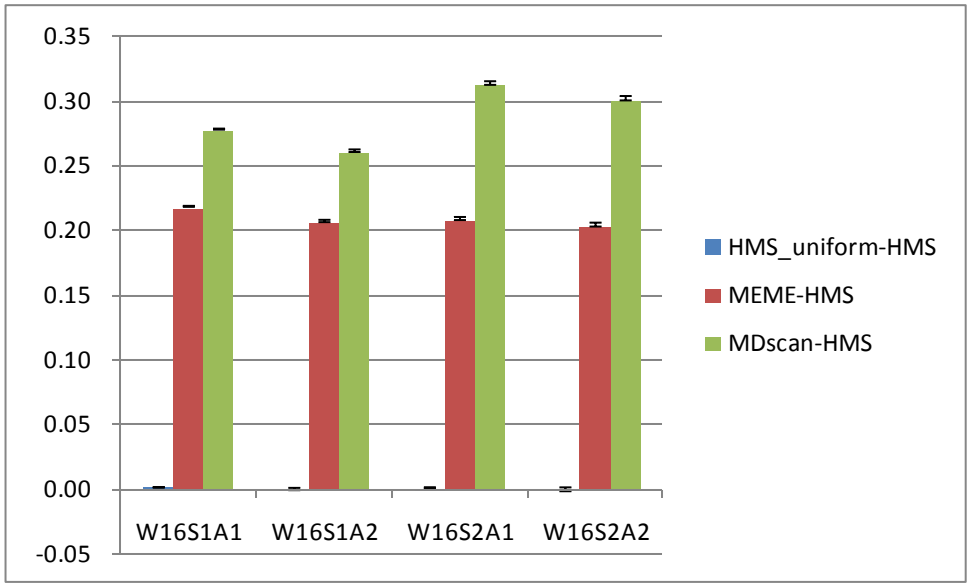
B. Independent, motif width = 16 bp.



C. Dependent, motif width = 8 bp.



D. Dependent, motif width = 16 bp.



We next compared performance on motif pattern prediction accuracy. The prediction accuracy is defined as the discrepancy between predicted and true Θ for each method and each dataset. See equation (3.11) for the expression for the discrepancy. To compare methods, within each simulation setting, we performed a paired *t*-test between

discrepancies obtained using HMS and that of a competing method (HMS_uniform, MEME and MDscan). Among the 100 datasets, we only considered the ones in which all methods successfully detected the right motif. Significantly smaller discrepancy (p-value <0.01) was observed in six out of eight simulation settings when comparing HMS to MEME, and in seven out of eight settings when comparing HMS to MDscan (Figure 3.1A, Figure 3.1B and Tables A3.2A in appendix). In addition, we found that adopting the informative prior for the proposed HMS method results in more accurate motif pattern prediction in all eight simulation settings than when using the uniform prior. (Tables A3.2A in appendix).

3.3.1.2 Inter-dependent motif models

We next conducted simulation studies to evaluate the performance of HMS when some positions within the motif showed inter-dependency. In our simulation, dependency was added to two pairs of positions in the 8 bp motif model and four pairs of positions in the 16 bp motif model. The joint distribution of the pairs was taken from the one predicted for position pair (1,2) in the E2F motif in Zhou and Liu (93) (as shown in Figure 2(b) in the original paper, reproduced in Table A3.3 in appendix).

In terms of motif detection, both versions of HMS achieved perfect results in five out of the eight simulation settings. MEME and MDscan achieved perfect results in four and two settings respectively. Furthermore, HMS and HMS_uniform reported higher detection rates compared to MDscan and MEME in all simulation settings. Our results also suggest that the HMS method assuming informative prior performed better than the HMS method assuming non-informative prior (Table A3.2B in appendix).

When comparing motif pattern prediction accuracy, paired *t*-tests showed that motif pattern prediction discrepancies were significantly smaller for HMS than MEME and MDscan in all testable simulation settings (MEME did not identify the correct motif in any dataset under two simulation settings; MDscan only identifies the correct motif in two out of 100 datasets under one simulation settings. Therefore no paired *t*-test is performed for those simulation settings). The performance was similar between the two versions of HMS (Figure 3.1C, Figure 3.1D and Table A3.2B in appendix).

3.3.2 Real data

To further evaluate the performance of HMS, we tested it along with MDscan and MEME on four real ChIP-Seq datasets. The first three datasets, namely NRSF (neuron-restrictive silencer factor) (4), STAT1 (signal transducer and activator of transcription protein 1) (68), and CTCF (CCCTC-binding factor) (67), are publically-available. The ER dataset, however, is newly generated for this study. The details of these four datasets can be found in Table A3.4A and appendix.

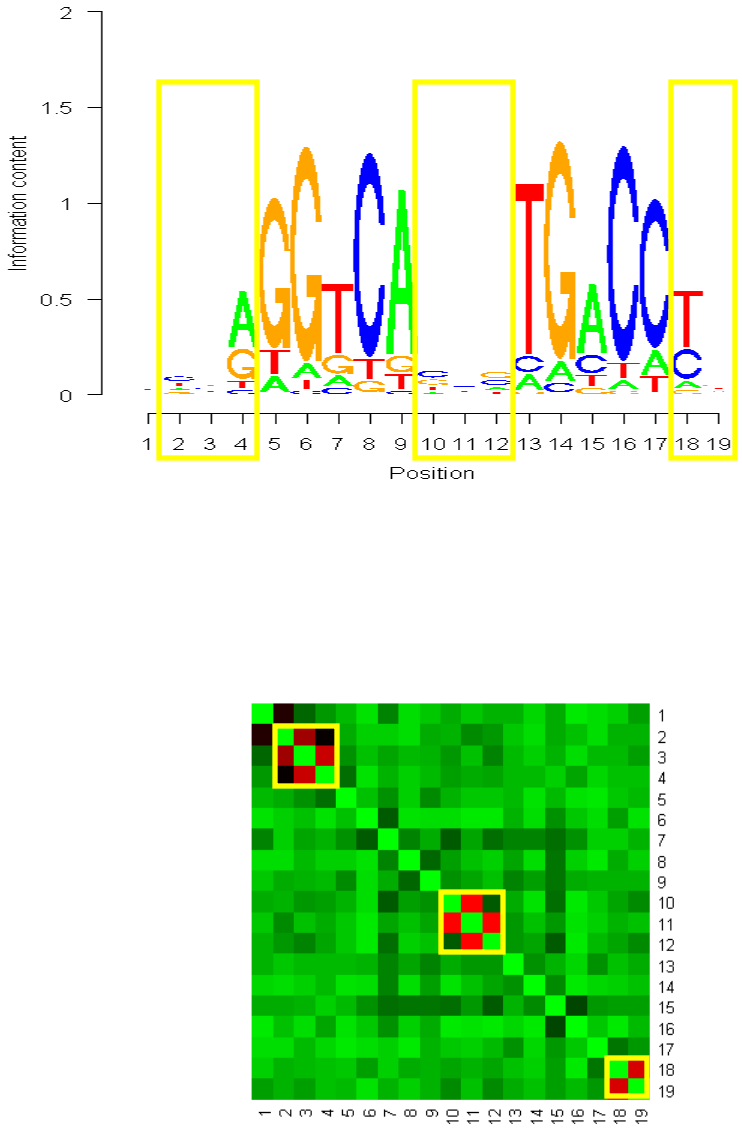
3.3.2.1 Intra-motif dependency

It is well known that some positions of TF binding motifs exert an inter-dependent effect on the binding affinities of TFs (87-90). However, due to the scarcity of the motifs identified for each TF, it is difficult to detect those dependent positions based solely on the limited motif sequence data. With the introduction of the ChIP-Seq technology, significantly more motif sequences can now be identified, which gives us unprecedented opportunity to identify dependent positions. Using the exhaustive search strategy we outlined in the methods section, we surveyed the four ChIP-Seq datasets used in this study: NRSF, STAT1, CTCF, and ER. The Hamming distance between two

probability vectors—$\{\theta_i, i = 1, \ldots, 16\}$ and $\{\theta_i \theta_j, i = 1, \ldots, 4, j = 1, \ldots, 4\}$ were presented in heatmaps (Figure 3.2 and Figure A3.3 in appendix). The two sets of probabilities of the 16 dinucleotides were estimated under the independent and dependent models respectively. Larger distance indicated higher dependency. Using the Hamming distance of 0.2 as the threshold, the number of dependent position pairs in the motif ranged from three to five in the four real datasets we studied (Table A3.5 in appendix). These pairs formed two triplets in NRSF and CTCF motifs, one triplet and one pair in the STAT1 motif and two triplets and one pair in the ER motif. In particular, we found that positions 14 and 15 in the CTCF motif show exceptionally strong dependency. The frequency of dinucleotides AC and GG in these positions were below what would be expected if they were independent. Similarly, the frequency for dinucleotides AG and GC exceeded expectations. The probability of discrepancy between independent and dependent motif models exceeded 0.1 in all four relevant cells in the four by four table (Table A3.6 in appendix).

Although our search strategy considers all pairs equally, we found that the strongest intra-motif dependency occurred at pairs of adjacent positions (Figure 3.2 and Figure A3.3 in appendix). All 16 dependent position pairs we identified in the four motifs were adjacent. This is not surprising given the strong dependency in neighboring positions of DNA sequences. We also found that strong intra-motif dependency often occurred in the so-called "gap" positions where the motif pattern appeared to be "weak" according to single-column motif model (e.g., positions 10, 11 and 11 and 12 in the ER motif).

Figure 3.2 Illustration of the unbiased exhaustive survey of all pairs of positions within the ER motif to assess the strength of their dependency. The differences in Hamming distance between the independent and dependent models are plotted in a heatmap. Larger differences (in dark red color) indicate higher dependency. Dependent triples: position 2, 3 and 4, position 10, 11 and 12. Dependent pairs: position 18 and 19. Dependent positions are illustrated in the box on the logo plot and the heatmap.



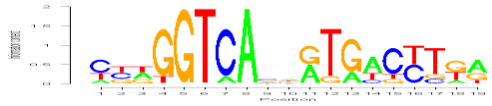
3.3.2.2 TFBS motif profile comparison

Since both HMS and MDscan were able to rapidly process tens of thousands of DNA sequences without sacrificing much computation time, we fed the entire set of ChIP-enriched regions into these two programs. In this comparison, we only used the top 500 sequences as input for MEME, since this program was not optimized to analyze large numbers of DNA sequences. Next, we applied MAST (96), a motif scanning software that is a companion to MEME, to scan the remaining sequences using the motif pattern identified by MEME. This is a commonly-used strategy in motif analysis (75). We also included motif patterns either from the literature (CTCF motif from Kim et al. (97)) or from MatBase (Genomatix, Software GmbH, Munich, Germany) for comparison. We used two different versions of HMS in our analysis: the default setting allowing dependency among positions in the motif and HMS_ind assumed all positions are independent. Informative prior for alignment variable A is used in both versions of HMS.

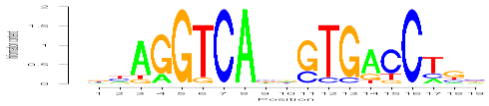
Figure 3.3 Comparison of ER motif patterns identified by different *de novo* motif-finding tools, as well as known motif patterns stored in the MatBase (Genomatix Software GmbH, Munich, Germany). A. Logo plots (52) of motifs identified by various motif-finding programs as well as the ones stored in the MatBase. The logo plots are generated using R package "seqLogo" (53). B. Comparison of motif enrichment in ChIP-Seq using cross validation. C. Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq. In order to obtain a smooth curve when plotting empirical FDR versus Chi-square test statistics, we applied kernel smoothing using an R function smooth.spline().

A.

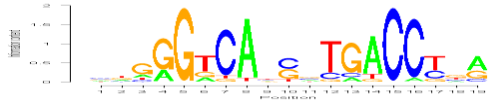
Genomatix V$ER01:



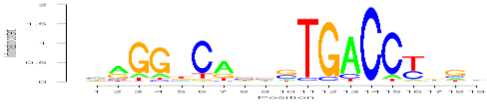
Genomatix V$ER02:



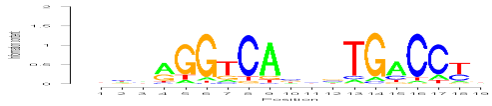
Genomatix V$ER03:



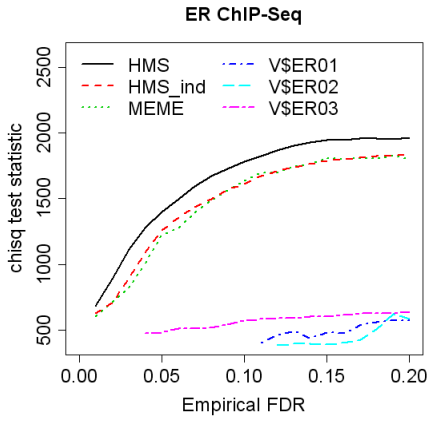
MEME:



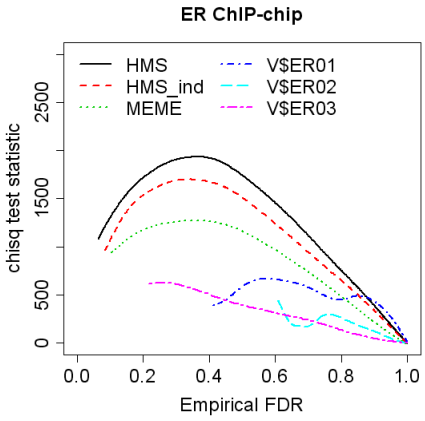
HMS:



B.



C.



Although the four TFs and their binding motifs were quite diverse, the motif pattern identification results were remarkably consistent. The results from the ER dataset are presented in Figure 3.3. Results from the three publicly available ChIP-Seq datasets can be found in Figure A3.4-A3.6 in appendix.

Figure 3.3A showed that *de novo* motif patterns identified by MEME and HMS from the ER ChIP-Seq dataset. Both patterns were similar to the ER motif stored in MatBase. However, the motif pattern identified by HMS was relatively less conserved (average information content: HMS: 0.64, MEME: 0.71, Genomatix V$ER01: 1.00, Genomatix V$ER02: 1.03, Genomatix V$ER03: 0.89) but more palindromic than the other motif patterns (Hamming distance between the two 6-mer half sites after one half site was converted to its reverse complement: HMS: 0.09, MEME: 2.57, Genomatix V$ER01: 4.00, Genomatix V$ER02: 2.18, Genomatix V$ER03: 2.53). We did not include MDscan in our comparison since MDscan was unable to consistently identify the consensus ER motif pattern. In Figure 3.3B, we plotted the Chi-square test statistics that measured the motif enrichment at different levels of the empirical FDR. Comparing AUC, we found that the motif patterns identified by MEME and HMS showed much higher AUC than the known motif patterns stored in MatBase. We believe that the dramatically increased number of binding sites identified by ChIP-Seq contributed to the refinement of the motif pattern. MEME and a simplified version of HMS (which used an independent mono-nucleotide model, referred as HMS_ind) exhibited a similar result. AUC for HMS, which allowed up to three-way interdependency, was 16.7% higher than MEME (Table A3.7 in appendix). The improvement is statistically significant when we repeated the cross-validation steps 100 times and compared the AUCs from HMS and

MEME using a paired t-test (p-value < 0.00001). We also compared the proportions of ChIP-enriched sequences that contain each of the ER motif patterns shown in Figure 3A. We found that, under the two empirical FDR levels (0.05 and 0.1), the proportion of motif pattern defined by HMS is higher than that from HMS_ind (by 12.95% and 8.07% respectively). Comparing HMS to MEME under these empirical FDR levels, the proportion of motif pattern defined by HMS again is higher (by 19.52% and 9.20% respectively). These differences are again significant (p-value < 0.00001) when verifying with paired t-test comparing results from 100 cross-validations. In addition, we found that proportions of motifs reported by HMS, HMS_ind and MEME are much higher than those found in the MatBase (Table A3.8 in appendix).

Among the other datasets (NRSF, STAT1, and CTCF), HMS and MEME consistently identified the consensus motif patterns in all trials. MDscan was able to consistently identify only the NRSF motif, but not the ones for the other two datasets. Again, we found that the motif patterns identified by these *de novo* motif-finding tools were more enriched than known motif patterns found in the literature or MatBase. Motif patterns defined by HMS consistently showed higher enrichment and resulted in higher AUC than MEME (Figures A3.4-A3.6, Table A3.7 in appendix). Motif patterns defined by HMS are consistently found in more ChIP-enriched sequences than those defined by HMS_ind and MEME at the same empirical FDR levels (Table A3.8 in appendix). The performance differences are significant except for the STAT1 motif.

3.3.2.3 Comparison to ChIP-chip data

In order to confirm that the higher enrichment of the motif identified by HMS on ChIP-Seq data was not platform-dependent, we compared an independent set of testing

and control sequences using ChIP-chip. Not only the technology is different, but also the cells and antibodies used. Detailed information about these datasets can be found in Table A3.4B and appendix.

Despite all the differences, we found that the ER motif pattern identified by HMS from ChIP-Seq data once again exhibited significantly higher enrichment than those of HMS_ind and MEME (Figure 3.3C): the improvements of AUC were 17.5%, and 57.4%, respectively (Table A3.7 in appendix). These differences are statistical significant (p-value < 0.00001). Similar plots and AUC comparisons performed on the other three datasets—NRSF, STAT1 and CTCF—showed comparable patterns (Figures A3.4-A3.6, Table A3.7 in appendix). These findings support that the motif pattern identified by HMS has a higher accuracy.

3.3.3 Computation Time

All computation was performed on Dell PowerEdge 1950 compute nodes with 2.83 GHz CPU processors and 8 GB RAM. To compare the computation time required for each algorithm, we selected the top 500, 1,000, 1,500, 2,000, up to 5,000 sequences identified from the NRSF ChIP-Seq data and fed them into the three motif-finding programs—MDscan, MEME, and HMS. We found MDscan to be the fastest, with HMS a close second. Computation time increased linearly with the number of sequences for MDscan and HMS; and both were much faster than MEME. The differences are quite dramatic. For real data, computation times for HMS ranged from 0.4 hours (NRSF data) to about 2.5 hours (CTCF data). However, since all parallel chains are independent, computation time can be reduced to one tenth if using a multi-processor computing

cluster. In contrast, MEME takes much longer; from 13 hours (NRSF data) to more than 23 days (CTCF data, job aborted after 23 days of running).

3.4 Discussion

The newly-emerged ChIP-Seq technology is capable of comprehensively revealing protein-DNA interacting sites across the entire genome with high resolution, which presents both opportunities and challenges for the identification of TFBS motif patterns. Increasing the number of input sequences allowed us to define TFBS motif patterns more accurately. However, most of the existing motif-finding programs such as MEME are not optimized to analyze the large number of input sequences that are generated from ChIP-Seq experiments. In this manuscript, we introduce HMS, a novel computational algorithm, specifically designed for TFBS motif discovery from ChIP-Seq data. It combines stochastic sampling with deterministic optimization in an iterative procedure. The assignment of sequences to these two treatments was dependent on the ranks of the ChIP-enrichment of those regions. This prioritized hybrid Monte Carlo strategy allows us to rapidly analyze tens of thousands of input sequences and produces an accurate estimate of the motif pattern. Our algorithm has the additional advantage of leveraging sequencing depth within each region to aid motif identification. Since the shape of sequencing depth is indicative of likely loci of the motif, using an informative prior gives HMS greater capability to identify weaker motifs than it could otherwise, a clear advancement.

In addition, using HMS we found that there is substantial intra-motif dependency among selected pairs of positions. We identified 16 highly significant position pairs within the NRSF, STAT1, CTCF and ER motifs. All of these position pairs are adjacent

to each other, some form triplets. In particular, we noticed a position pair (14 and 15) in the CTCF motif that displays exceptionally strong dependency in which dinucleotides AG and GC are far more frequent then AC and GG at these two positions. Using both simulated data and real data, we showed that incorporating dependent positions in a motif model offers further improvement in detecting and characterizing the underlying TF binding motif patterns.

Currently, most *de novo* motif searches on sequences identified by ChIP-Seq are conducted on a subset of all available sequences. This is because searching through the full set of thousands, or even tens of thousands, of input sequences using existing motif-finding tools is extremely time-consuming. Our simulation study showed that this strategy, while convenient, has increased the likelihood of missing the true motif patterns. Further, the probability matrix Θ inferred with this strategy are often less accurate. In contrast, HMS allows us to analyze the full set of input sequences within only a fraction of the computational time required for existing *de novo* motif-finding tools like MEME. In this study, stochastic search was performed on the top 10% of all sequences. This proportion is adjustable by users. We have experimented increasing or decreasing the 10% cutoff and found that these changes made little difference in the performance of HMS. When applied to multiple real ChIP-Seq datasets, we found that the motif patterns identified by HMS tend to be more enriched than motifs identified by other methods. Remarkably, when comparing the same motif patterns identified from ChIP-Seq data to enriched regions identified from independent ChIP-chip experiments for the same TF, even with different cell types or different antibodies or both, we still found that motif patterns identified by HMS showed higher enrichment in the ChIP-enriched regions

relative to random control sequences. This finding suggests that the motif patterns identified by HMS are closer to the underlying motif pattern recognized by the TF.

In this study, we utilized ChIP-enrichment of the peaks to rank order all input sequences, believing that ChIP-enrichment is positively correlated with the motif abundance. However, there are many potential reasons, both biological and technical, that a particular region is sequenced more deeply. These include the availability of the antibody's epitope during the immunoprecipitation step, conformational changes on the TF, abnormality in the cell line such as aneuploidy, bias introduced during the sequencing library construction, nucleotide-induced sequencibility bias (such as GC content) and bias related to alignment (repeat regions, various polymorphisms). These complications will reduce the correlation between ChIP-enrichment and sequencing depth. We believe advanced models that consider these factors will further improve the performance of HMS. Another potential enhancement would be to model the protein-DNA binding affinity indicated by read density using thermodynamic models (98).

In this study, if the motif width is unknown, we run HMS with every possible width within the range specified by the user and report all significant motif patterns. One possible improvement to this step would be to allow motif width w to vary during iterations (99). For example, we may add a Metropolis step, with equal probability of adding or removing one base at one end of the motif, and test whether the new motif pattern provides a better fit with the data. Another possible area for improvement concerns multiple binding sites. Currently, HMS is only designed to search for the primary binding site (i.e., the binding motif of the regulatory protein being ChIP'ed).

However, we can also use HMS to identify secondary binding sites by masking the first motif identified and re-running HMS on the masked sequences.

In summary, we showed that ChIP-Seq data can significantly increase our ability to discover and refine TFBS motif patterns. However, new computational tools are needed in order to efficiently and thoroughly handle the ChIP-Seq data, as well as to exploit the various advantages of ChIP-Seq technology. The development of the highly scalable HMS algorithm represents an early attempt. With significant improvement in both accuracy and computation speed, we believe that HMS will be of broad interest to researchers conducting ChIP-Seq experiments and has the potential to accelerate discovery in biomedical research.

3.5 Appendix

3.5.1 URLs of data used in this study

ChIP-Seq:

NRSF:

http://www.illumina.com/downloads/Illumina_ChIPSeq_Demo_Data_Johnson_Science_2007.zip

STAT1: http://www.bcgsc.ca/data/chipseq

CTCF: http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx

ER:

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM470418

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM470419

ChIP-chip:

NRSF:

http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds&term=GSE8489[Accession]&cmd=search

STAT1: http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE2714

CTCF: http://licr-renlab.ucsd.edu/download.html

ER:

http://research4.dfci.harvard.edu/brownlab//datasets/index.php?dir=ER_MCF7_whole_human_genome/

3.5.2 HPeak Software

HPeak (95) is a hidden Markov model (HMM)-based algorithm for analyzing ChIP-Seq data. The goal of HPeak is to partition the genome into segments that are either ChIP-enriched or non-enriched such that the enriched portion of the genome is much more likely to harbor protein-DNA interaction sites. The input data is a collection of sequencing reads that have been aligned to the reference genome uniquely. HPeak first partitions the entire genome into small bins of fixed length (e.g., 25 bps) and evaluates the distribution of ChIP DNA fragments in these bins throughout the genome. Next HPeak applies a two-state HMM on the sequencing depth profile to identify stretches of ChIP-enriched bins from the background. HPeak uses two different probability distributions, the generalized Poisson (GP) distribution (100) and the zero inflated Poisson (ZIP) distribution (101) to model the numbers of sequencing reads that overlap with ChIP-enriched and non-enriched bins respectively. Both these distributions are modified from the standard Poisson distribution to fit data where there is serious over or under dispersion or there are a large proportion of extra zeros, cases we often observe in

genome-wide sequencing read coverage profiles. Using a user-specified posterior probability threshold, HPeak then identifies stretches of bins from the HMM that show significant enrichment of sequencing read counts. Each set of bins is defined as a peak. In addition to its genomic location and the length of the peak, HPeak also reports the location of the highest sequencing depth within the peak, the actual maximum sequencing depth at that location, and the posterior probability of these bins being ChIP-enriched. Because such probability reflects the significance of these peaks, one can rank all peaks predicted by HPeak using these probabilities. HPeak software is freely available from the website http://www.sph.umich.edu/csg/qin/HPeak/.

3.5.3 Estrogen Receptor ChIP-Seq Experiment on MCF7 Cells

MCF-7 cells were grown in RPMI media supplemented with 10% FBS to 50% confluence. The cells were then hormone-starved for three days prior to the treatment with 10 nM β-estradiol or vehicle control for 45 minutes. The cells were then harvested for ChIP analysis as previously described using an antibody against the estrogen receptor (ER)-alpha (sc-543x, Santa Cruz Biotechnology Inc, Santa Cruz, CA) or the IgG control. Briefly, cultured cells near 90% confluence were crosslinked with 1% formaldehyde for 10 minutes and the crosslinking was inactivated by 0.125 M glycine for 5 minutes at room temperature (RT). The cells were then rinsed with cold 1X PBS twice and scraped off in 1X PBS + protease inhibitor (PI). Cells were pelleted and resuspended in cell lysis buffer plus PI for 10 minutes. Nuclei pellets were spun at 5,000 rpm for 5 minutes, resuspended in nuclear lysis buffer, and then incubated for 10 minutes. Chromatin was sonicated to an average length of 500 bp with an Ultrasonic Processor Sonicator 3000 (Misonix Inc, Farmingdale, NY) and then centrifuged at 14,000 rpm for 10 minutes to

remove the debris. Supernatants containing chromatin fragments were incubated with agarose/protein A or G beads (Millipore. Billerica, MA) for 15 minutes and centrifuged at 5,000 rpm for 5 minutes to remove the nonspecific binding. To immunoprecipitate protein/chromatin complexes, the supernatants were incubated with 3-5 μg of antibody or IgG overnight, then added 50 μl of agarose/protein A or G beads and incubated for another hour. Beads were washed twice with 1X dialysis buffer and four times with IP buffer. The antibody/protein/DNA complexes were eluted with 150 μL IP elution buffer twice. To reverse the crosslinks, the complexes were incubated in elution buffer plus 10 μg RNase A and 0.3 M NaCl at 67°C for four hours. DNA/proteins were precipitated with ethanol, air-dried, and dissolved in 100 μL of TE. Proteins were then digested by proteinase K at 45°C for one hour and DNA was purified with a QIAGEN PCR purification column and eluted with 30 μL EB buffer.

The ChIP-enriched DNA was evaluated for significant enrichment of positive control genes and then subjected to ChIP-Seq sample preparation following the manufacturer's protocols (Illumina Inc, San Diego, CA). Briefly, the ends of ChIP-enriched DNA or control DNA (~10 ng) was first repaired by T4 DNA polymerase, T4 PNK, and Klenow DNA polymerase at 20°C for 30 minutes. An "A" base was added to the 3' end of the blunt phosphorylated DNA fragments using Klenow exo at 37°C for 30 minutes. Adapters were then ligated to the ends of the DNA fragments by DNA ligase at RT for 15 minutes. DNA fragments were separated on 2% gel at 100V for 1 hour, the 200±25bp band was excised from the gel, and the DNA was extracted by QIAGEN gel extraction kit. Gel-extracted DNA was amplified by PCR reaction for 16 cycles and quality assured using Bioanalyzer (Agilent Technologies, Santa Clara, CA). ChIP-

Sequencing was performed using the Illumina Genome Analyzer according to standard manufacturer's procedures. The raw sequencing image data were analyzed by the Illumina analysis pipeline, aligned to the unmasked human reference genome (NCBI v36, hg18) using Eland software (Illumina Inc, San Diego, CA) to generate sequence reads of 35 bps.

3.5.4 Simulation scheme for studying intra-motif dependency

As described in the Methods section of this chapter, for a pair of positions within the motif, we use the Hamming distance between two sets of estimated dinucleotide frequencies based on two competing probability models (16-component multinomial distribution or the product of two four-component multinomial distributions) to gauge whether the two positions are dependent. To select a reasonable cutoff, we conducted a simulation study to estimate the null distribution for such Hamming distances. We considered five levels of nucleotide conservation, with information content ranging from 0.29 to 1.76 (0.29, 0.64, 1.15, 1.42 and 1.76). There are in total 15 different combinations of these information contents. For each combination, we specify two four-component multinomial distributions that match the two information content levels. One thousand nucleotides were drawn from each of the two multinomial distributions independently. We choose the large number to reflect the fact that typically large amounts of motifs were identified from ChIP-Seq experiments. The Hamming distances were calculated using formula (3.4) in this chapter:

$$d_{12} = \sum_{x=1}^{4}\sum_{y=1}^{4} |\hat{\eta}_{xy}(r_1, r_2) - \hat{\eta}_x(r_1)\hat{\eta}_y(r_2)|$$

We simulated one million position pairs using the above procedure in order to obtain an accurate null distribution of the background Hamming distance. The histograms of these Hamming distances were shown in Figure A3.7. From these plots we found that strong dependency (large Hamming distance) tends to occur between a pair of positions in which each position itself is weakly conserved.

3.5.5 Tables in appendix

Table A3.1A Four motif models for two motif widths and two motif strengths used in the simulation study.

Motif Consensus (Width)	Motif information content	
	S1 (high)	S2 (low)
GACTACCA (W8)	1.42	0.93
AGGATCTAATGATCCT (W16)	1.42	0.93

Table A3.1B Two motif abundances scheme used in the simulation study.

Expected copies of motif segments	Motif abundance	
	A1 (high)	A2 (low)
Top 25% sequences	0.9	0.6
25% - 50% sequences	0.7	0.4
50% - 75% sequences	0.5	0.2
Last 25% sequences	0.3	0
Total expected motif segments	1,800	900

Table A3.2A Performance comparison on simulated data assuming all positions within the motif are independent.

Simulation setting[1]	Times_found[2]				Difference compared to HMS[3]		
	HMS	HMS_uniform	MEME	MDscan	HMS_uniform	MEME	MDscan
W8S1A1	100	100	100	100	0.01**	0	0.11**
W8S1A2	100	100	100	100	0.00**	0	0.09**
W8S2A1	100	100	71	33	0	0.07**	0.11**
W8S2A2	100	100	78	96	0.01	0.09**	0.02
W16S1A1	100	100	100	100	0	0.02**	0.12**
W16S1A2	100	100	100	100	0.00**	0.02**	0.10**
W16S2A1	100	100	100	54	0.00*	0.02**	0.21**
W16S2A2	100	100	100	91	0.00*	0.02**	0.20**

Table A3.2B Performance comparison on simulated data assuming some positions within the motif are dependent.

Simulation setting[1]	Times_found[2]				Difference compared to HMS[3]		
	HMS	HMS_uniform	MEME	MDscan	HMS_uniform	MEME	MDscan
W8S1A1	98	98	63	96	0	0.10**	0.11**
W8S1A2	100	100	85	93	0	0.09**	0.10**
W8S2A1	92	89	0	2	-0.01#	NA	0.09#
W8S2A2	73	69	0	61	0.03	NA	0.09**
W16S1A1	100	100	100	100	0	0.22**	0.28**
W16S1A2	100	100	100	100	0	0.21**	0.26**
W16S2A1	100	100	100	40	0	0.21**	0.31**
W16S2A2	100	100	100	83	0	0.20**	0.30**

[1]Each simulation setting is a combination of motif width (W), information content (S) and motif abundance (A). The scheme is similar to Liu et al, (2001) (54) and described in Table A3.1.

[2]"Times found" indicates among the 100 simulated dataset, how many times the correct motif is identified by the motif-finding algorithm.

[3]Difference referrers to the difference between two average Hamming distances h and h_{HMS} in which h measures average discrepancy (per base) between probability matrix Θ and its prediction denoted as $\hat{\Theta}$: $h = \frac{1}{w}\sum_{i=1}^{4}\sum_{j=1}^{w}|\theta_{ij} - \hat{\theta}_{ij}|$ (Formula (3.11)). h_{HMS} measures average discrepancy (per base) between probability matrix Θ and $\hat{\Theta}$ predicted by HMS method. We use * to indicate a p-value in paired t-test between 0.01 and 0.05 and ** to indicate a p-value in paired t-test less than 0.01. [#] indicates that p-value is not available due to lack of sample size (in simulation setting W8S2A1, MDscan can detect correct motif only in two out 100 simulated data sets.)

Table A3.3 The joint distribution of dinucleotides in two dependent positions. The probabilities in this multinomial distribution are taken from the one predicted for position pair (1,2) in the E2F motif in Zhou and Liu (93) (Figure 2(b) in their paper).

First Base	Second Base			
	A	C	G	T
A	0	0	0.19	0
C	0	0	0.16	0.06
G	0	0.09	0	0
T	0	0.44	0.06	0

Table A3.4A Information on the four real ChIP-Seq datasets

TF	Cell type	Antibody	# of peaks	Coverage	Reference
NRSF	Jurkat T cell	Monoclonal antibody 12C11	4,982	1.4 MB	Johnson et al. (2007)
STAT1	HeLa S3 cell	Rabbit polyclonal antibody	27,470	8.1 MB	Robertson et al. (2007)
CTCF	CD4+ T cell	Upstate 07-729	22,159	7.4 MB	Barski et al. (2007)
ER	MCF7 cell	ER α (HC-20)	10,072	2.5 MB	Hu et al. (2010)

Table A3.4B Information on the four real ChIP-chip datasets

TF	Cell type	Antibody	# of peaks	Coverage	Reference
NRSF	Jurkat T cell	Monoclonal antibody	8,819	12.2 MB	Johnson et al. (2007)
STAT1	HeLa S3 cell	α p91 (C-24) rabbit polyclonal antibody	3,701	4.7 MB	Euskirchen et al. (2007)
CTCF	IMR90 and U937 cell	Mixture of 9 monoclonal antibodies	13,804	12.1 MB	Kim et al. (2005)
ER	MCF7 cell	ER α (HC-20)	10,901	11.1 MB	Carroll et al. (2005)

Table A3.5. Dependent positions identified in four motifs from ChIP-Seq data.

Motif	Top dependent pairs	Hamming distance
NRSF width = 21 bp	(18-19)	0.3308
Dependent positions:	(3-4)	0.3297
[(2-3)(3-4)] [(18-19)(19-20)]	(19-20)	0.3196
	(2-3)	0.2629
CTCF width = 24 bp	(14-15)	0.4635
Dependent positions:	(16-17)	0.3358
[(13-14)(14-15)] [(16-17)(17-18)]	(17-18)	0.2720
	(13-14)	0.2664
STAT11 width = 19 bp	(7-8)	0.3052
Dependent positions:	(6-7)	0.2853
[(6-7)(7-8)] (13-14)	(13-14)	0.2008
ER width = 19 bp	(11-12)	0.2562
Dependent positions:	(10-11)	0.2553
[(2-3)(3-4)] [(10-11)(11-12)] (18-19)	(18-19)	0.2327
	(3-4)	0.2260
	(2-3)	0.2070

Table A3.6. The probability mass function of 16-component multinomial distribution (dependent) of position 14 and position 15 in the CTCF motif, and the probability mass function of 16-component multinomial distribution (independent) of position 14 and position 15 in CTCF motif (the outer product of the probability mass function of two independent four-component multinomial distributions).

	Position 15				Marginal distribution
Position 14	A	C	G	T	
A	0.03 (0.04)	**0.14 (0.25)**	**0.28 (0.16)**	0.03 (0.03)	0.48
C	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01
G	0.05 (0.04)	**0.34 (0.24)**	**0.06 (0.17)**	0.03 (0.03)	0.48
T	0.00 (0.00)	0.02 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02
Marginal distribution	0.08	0.51	0.34	0.06	1

Table A3.7. Comparison of motif enrichment among motif patterns identified by different *de novo* motif finding tools as well as known motif patterns described in the literature or stored in the MatBase (Genomatix Software GmBH, Munich, Germany).

Area under the curve[1]	ChIP-Seq[2]				ChIP-chip[3]			
	NRSF	STAT1	CTCF	ER	NRSF	STAT1	CTCF	ER
HMS	**258.62****	439.68	**2312.87****	**327.99****	**163.08****	**94.23****	**1102.02****	**1216.94****
HMS_ind	254.73**	388.99	2198.81**	290.30	161.98**	83.75	1001.35**	1035.42**
MEME	242.83	**440.33**	2076.76	281.14	148.76	82.95	908.35	773.30
MDscan	240.16	--	--	--	143.25	--	--	--
Genomatix V$NRSF	210.38	--	--	--	36.38	--	--	--
Genomatix V$STAT01	--	77.39	--	--	--	9.69	--	--
Genomatix V$STAT03	--	191.70	--	--	--	28.96	--	--
Kim07_CTCF	--	--	1225.97	--	--	--	509.62	--
Genomatix V$ER01	--	--	--	35.29	--	--	--	292.43
Genomatix V$ER02	--	--	--	35.05	--	--	--	71.48
Genomatix V$ER03	--	--	--	87.40	--	--	--	245.37

[1]Area under the curve (AUC) in the empirical FDR versus Chi-square test statistics plot (Figure 3.3B and C, Figures A3.4-A3.6 B and C). Values in bold indicate the best performance in that column. Five cross-validations were performed on each dataset using each of the four motif finding algorithms. In addition, we conducted cross-validation 100 times and compared the AUCs obtained from two different method using a paired t-test to assess whether the performance difference we observed in statistical significant. We use ** indicates an empirical p-value less than 0.01.

[2]The empirical FDRs for NRSF, STAT1, CTCF and ER ChIP-Seq data sets all range from 0 to 0.2.

[3]The empirical FDRs for NRSF and CTCF ChIP-chip data sets range from 0 to 0.2. The empirical FDRs for STAT1 and ER ChIP-chip data sets range from 0 to 1.

Table A3.8. Comparison of motif enrichment among the three motif finding tools and known motif patterns stored in the MatBase (Genomatix GmBH, Munich, Germany).[1]

ChIP-Seq	NRSF		STAT1		CTCF		ER	
Empirical FDR[2]	0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1
HMS	**49.30**	**51.41**	**10.84**	**25.29**	**79.05**	**84.45**	**29.39**	**38.57**
HMS_ind	48.25	51.09	10.12	22.10	75.91	83.10	26.02	35.69
MEME	46.63	49.55	10.28	24.08	72.43	81.16	24.59	35.32
MDscan	45.21	49.11	--	--	--	--	--	--
Genomatix V$NRSF	41.09	42.52	--	--	--	--	--	--
Genmoatix V$STAT01	--	--	--	--	--	--	--	--
Genomatix V$STAT03	--	--	10.09	12.16	--	--	--	--
Kim07_CTCF	--	--	--	--	56.68	62.01	--	--
Genomatix V$ER01	--	--	--	--	--	--	--	--
Genomatix V$ER02	--	--	--	--	--	--	--	--
Genomatix V$ER03	--	--	--	--	--	--	10.62	14.56

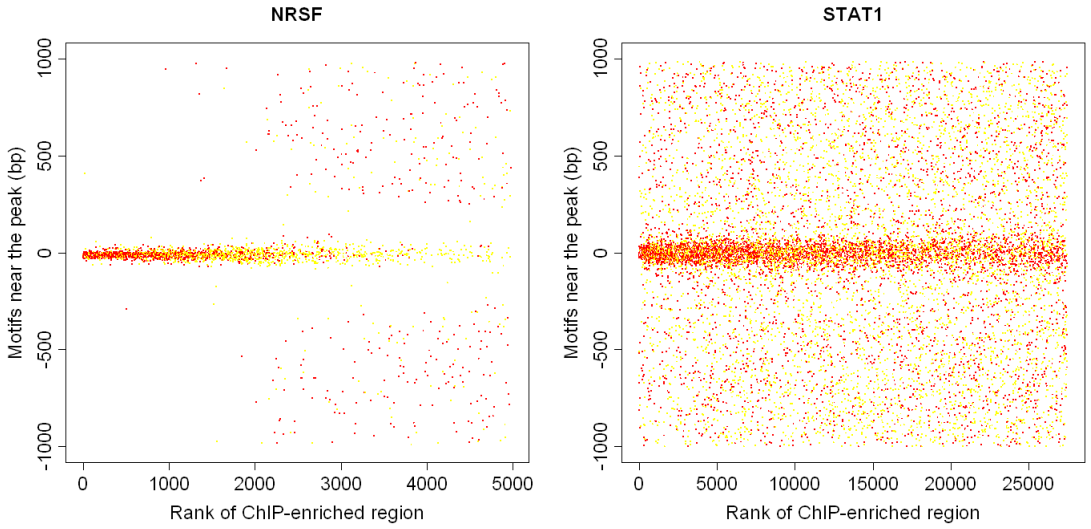
[1]Values in the table are percentages of ChIP-enriched sequences that contain the specific motif pattern. Values in bold indicate the best performance in that column.

[2]The empirical FDR is estimated by dividing the number of control sequences that contain the motif by the number of testing sequences that contain the motif.

3.5.6 Figures in appendix

Figure A3.1 Rank order of all ChIP-enriched regions versus location of motifs. Zero in the y-axis indicates the location of highest sequencing depth in the ChIP-enriched regions obtained from HPeak program. For each position $a_j = l$ within a ChIP-enriched region R_j, we calculate a motif score defined as equation (3.1), measuring the similarity between the DNA sequence of length w (motif length, assumed known) starts from the current location and the known motif pattern represented by PSWM. Higher scores indicates better match. We record the position with highest motif score for each ChIP-enriched region R_j. For each dot, the x-axis represents the rank of ChIP-enriched region (from the highest to the lowest), and the y-axis represents the physical position of the most likely motif location in each ChIP-enriched region. The red dots indicate the motifs with score above the first quantile, and the yellow dots indicate the motifs with score between the first quantile and median. A. NRSF, B. STAT1, C. CTCF, D. ER.

A. B.



C. D.

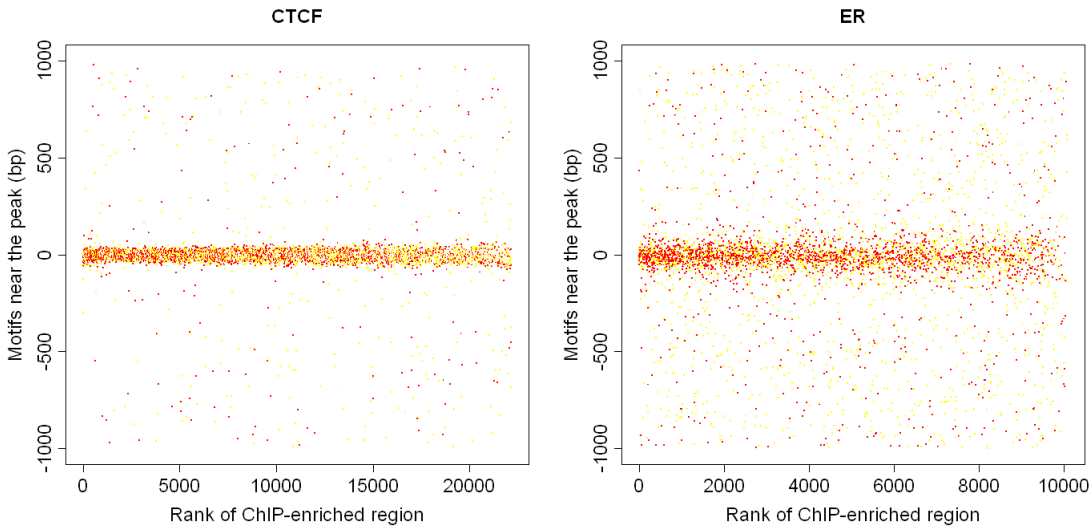


Figure A3.2 Illustration of the informative prior distribution of motif start locations. The prior probabilities (solid black line) are proportional to a discretized Student's *t* distribution with three degrees of freedom (with standard error = 1.73) and rescaled such that the prior probabilities form a step function with a fixed step-size (25 bp in this study). The solid red line represents the probability density function of shifted and rescaled Student's *t* distribution with three degrees of freedom.

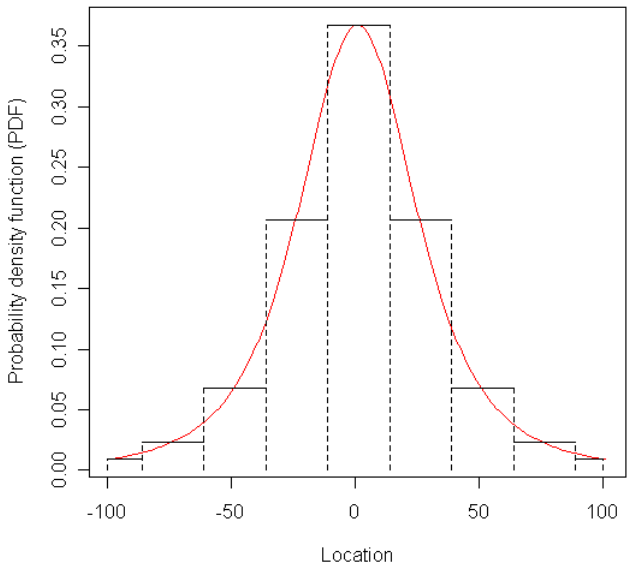
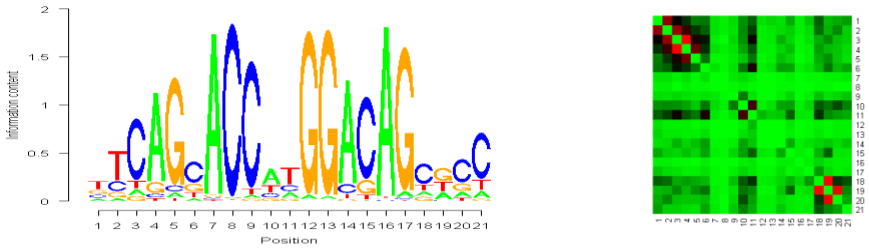
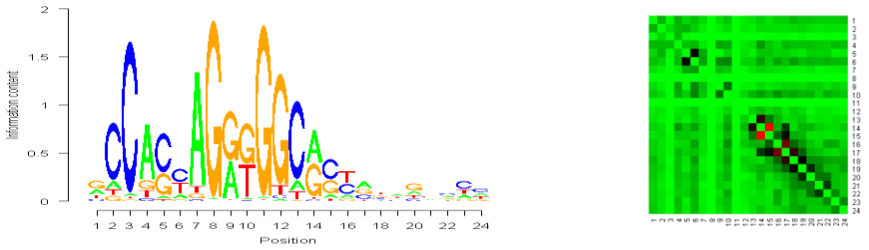


Figure A3.3 Illustration of the unbiased exhaustive survey of all pairs of positions within the motif. Larger differences (in darker color) indicate higher dependency.

NRSF: (2,3,4) (18,19,20)



CTCF: (13,14,15) (16,17,18)



STAT1: (6,7,8) (13,14)

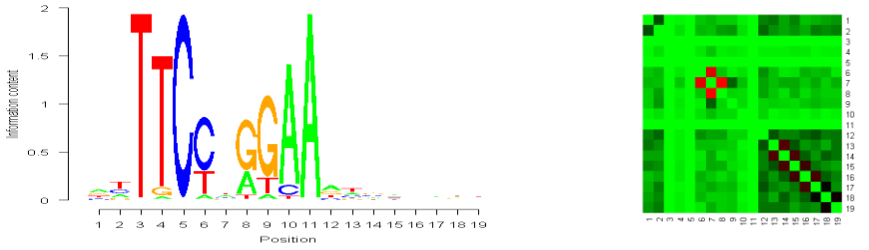
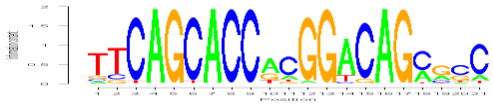


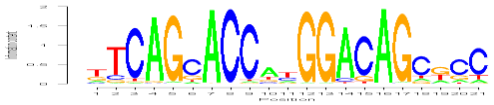
Figure A3.4 Comparison of NRSF motif patterns identified by different de novo motif finding tools as well as the known NRSF motif found in the MatBase (Genomatix Software GmBH, Munich, Germany). A. Logo plots of motifs identified by various motif finding programs as well as the NRSF motif stored in the MatBase. B. Comparison of motif enrichment in ChIP-Seq using cross validation. C. Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq.

A.

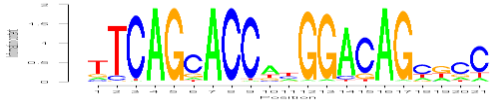
Genomatix:



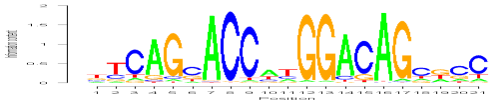
MEME:



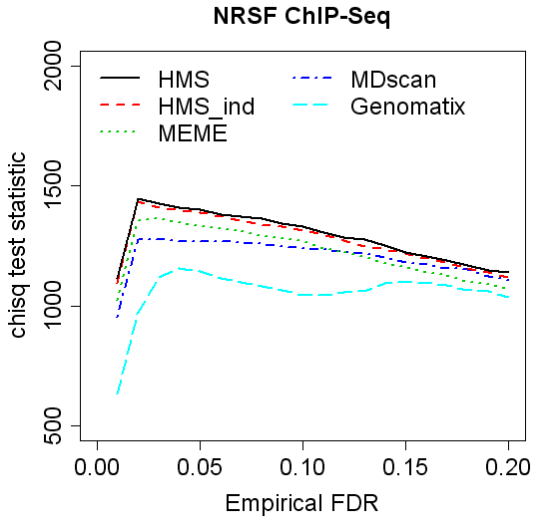
MDscan:



HMS:



B.



C.

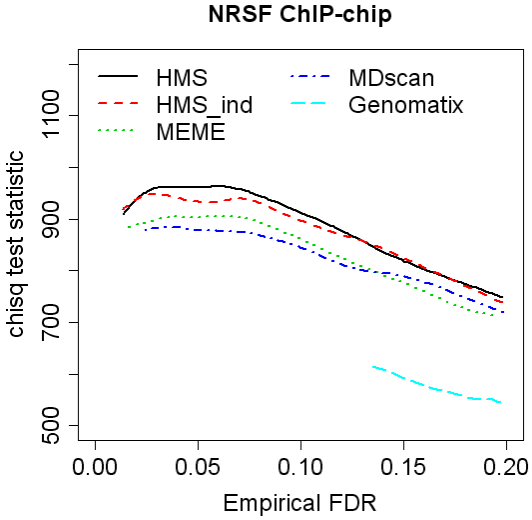
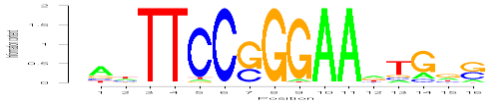


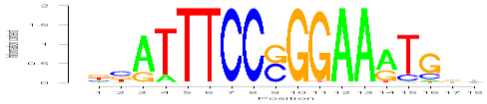
Figure A3.5 Comparison of STAT1 motif patterns identified by different de novo motif finding tools as well as known STAT motif patterns stored in the MatBase (Genomatix Software GmBH, Munich, Germany). A. Logo plots of motifs identified by various motif finding programs as well as the STAT motifs stored in the MatBase. B. Comparison of motif enrichment in ChIP-Seq using cross validation. C. Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq. Note: the x axis in STAT1 ChIP-chip figure is from 0 to 1.0 instead of the usual range of 0 to 0.2 due to its high empirical FDR.

A.

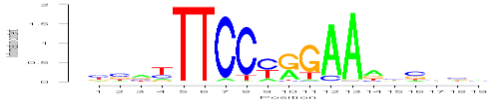
Genomatix V$STAT01:



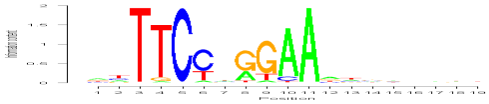
Genomatix V$STAT03:



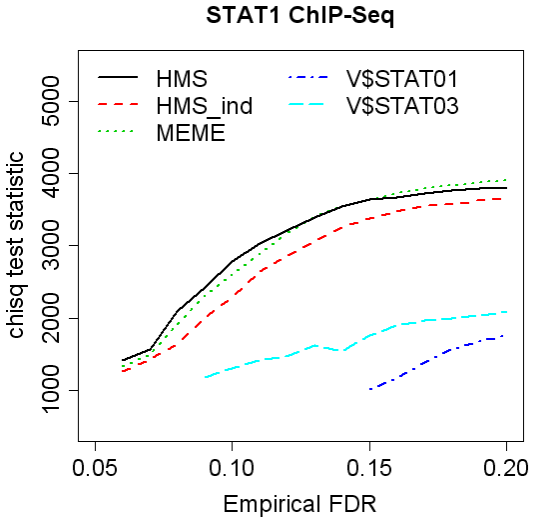
MEME:



HMS:



B.



C.

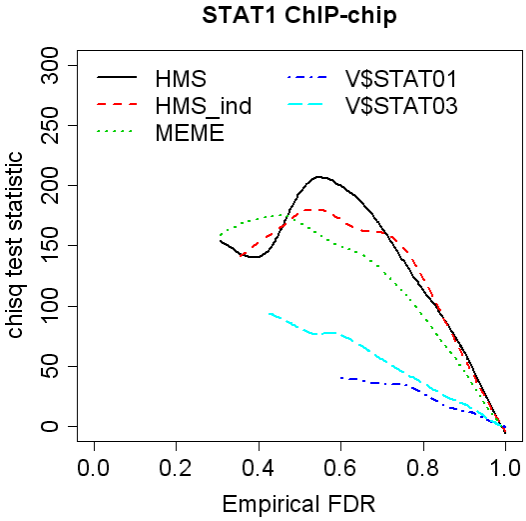
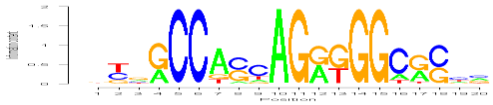


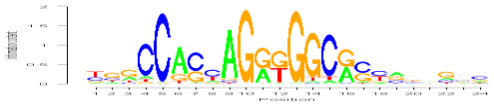
Figure A3.6 Comparison of CTCF motif patterns identified by different de novo motif finding tools as well as a known motif pattern found in Kim et al. (2007). A. Logo plots of motifs identified by various motif finding programs as well as the one found in Kim et al. (2007). B. Comparison of motif enrichment in ChIP-Seq using cross validation. C. Comparison of motif enrichment in ChIP-chip data using motif patterns identified in ChIP-Seq.

A.

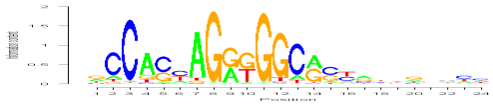
Kim07_CTCF:



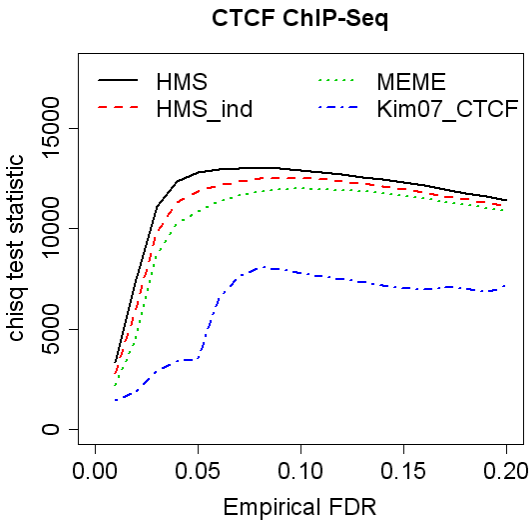
MEME:



HMS:



B.



C.

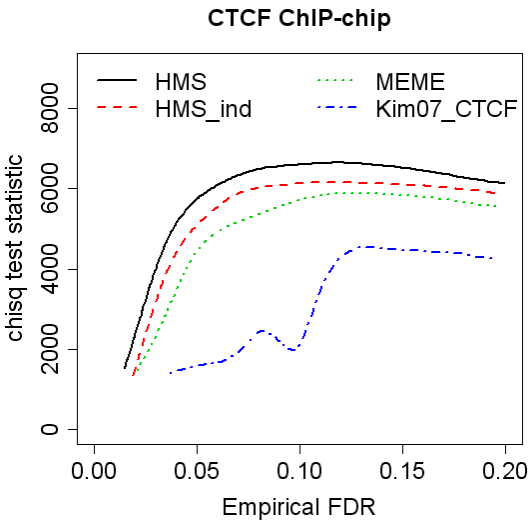
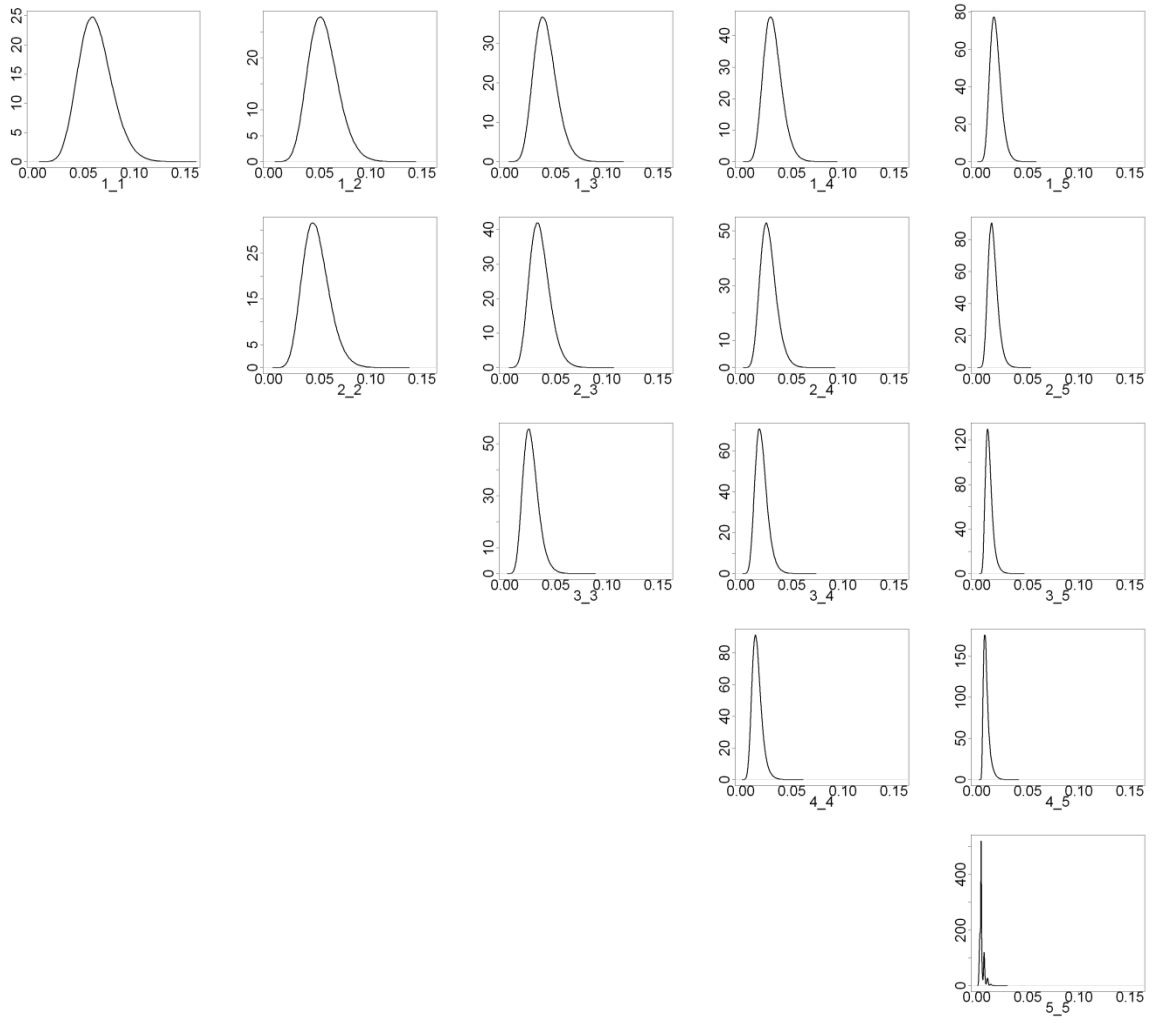


Figure A3.7 The histogram of 15 empirical distributions of Hamming distance. The label $i\_j, 1 \leq i \leq j \leq 5$ shown under each of the 15 empirical distributions of Hamming distance indicates the 15 combinations of information content. Small i, j represent low information content, while large i, j represent high information content.



CHAPTER 4

Spatial Poisson regression model for gene expression using RNA-Seq data

4.1 Introduction

The transcriptome is the complete set of transcripts in a cell under any given developmental stage or physiological condition. Comprehensively cataloging all the components in the transcriptome is a grand challenge in the post-genome era. In the past decade, microarray technology has played a prominent role in advancing our understanding of transcriptome complexity. Microarray is a hybridization-based technology that incubates fluorescently-labeled cDNA with custom-made microarray or commercial GeneChips. Microarray allows scientists to simultaneously monitor the expression of almost all the genes in the genome, and along with a steady reduction in processing costs, led to its wide spread application. Despite its overwhelming success, microarray technology has its limitations. First, design probes on the microarray requires knowledge of the genome sequence, hence novel transcripts will be missed. In addition, cross-hybridization, background signal and saturation result in a reduction of its dynamic range.

Sequencing-based approaches to measure gene expression have the potential to overcome these limitations. The ultra-high-throughput next generation sequencing (also known as massively parallel sequencing) technologies capable of producing millions of sequence reads are making the transition from development to widespread application rapidly. These technologies are able to dramatically increase the throughput in DNA sequencing compared to conventional Sanger technology and at much lower cost. An array of studies have been published that successfully apply these new sequencing technologies to measure mRNA expression levels on cells from various species including *Saccharomyces cerevisias*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, mouse and human cells (102-108). In RNA-Seq experiments, a population of RNA is converted to a library of cDNA fragments with adaptors attached to one end. Each molecule, after amplification, is then sequenced using one of the next generation sequencing technologies. Following sequencing, the resulting reads are aligned to either the reference genome or known transcripts to produce a genome-scale transcriptional profile.

While sequencing costs have significantly declined with the advent of the new technologies, the amount of data the new platforms produce is skyrocketing, thereby producing an analytical bottleneck. To match the advancement provided by the new sequencing technologies, significant attention and effort needs to be directed to the statistics and bioinformatics front. Sophisticated and tailor-made data analysis strategies are needed in order to fully realizing the power of the new sequencing technologies.

In order to quantify and compare transcriptions between different genes or between different experiments, read data generated by RNA-Seq needs to be properly normalized. Several methods for transcript quantification have been proposed in the

literature. Mortazavi et al. (2008) (104) proposed measuring the transcript levels in reads per kilobase of exon per million mapped reads (RPKM), which takes into consideration RNA length and the total read number in the measurement. The RPKM measure is essentially an averaging estimate of the transcript's expression intensity assuming an underlying Poisson or binomial model. Similar approaches have been used by other authors, for example, Bullard et al. (2009) (109) and Marioni et al. (2009) (107). All these methods utilize reads data at either the gene level or at the exon level. The counts of the bases demonstrate fairly large base-specific variations and between-base correlations (see preliminary analysis of a real RNA-Seq dataset in appendix), which we believe are attributed partly to the RNA-Seq technology and partly to some transcription mechanisms. Most current methods, including the RPKM measure and the method proposed by Jiang and Wong (2010) (110), do not take into consideration the base-level variations and correlations. As a result, the estimation of transcript expression will become less efficient, especially for low-expressed isoforms. In addition, it will compromise the comparison of transcript expressions between genes or between different experiments and hinder the discovery of novel transcript activities.

Because between-base correlations depend on the relative distances between bases, they are referred to as one-dimensional spatial correlations or in short spatial correlations in this study. The presence of location-specific variations along with spatial correlation is a characteristic of many spatial data sets generated in geostatistics, spatial epidemiology, and image processing and has been studied in the literature of spatial statistics. In this study, we apply and extend the ideas, models, and methodologies rooted in spatial statistics to model and analyze RNA-Seq data.

4.2 Methods

4.2.1 Spatial Poisson regression model

We consider RNA-Seq data at the single exon level in this study. Observed sequencing depth often demonstrates substantial variation and spatial correlation (see preliminary analysis of a real RNA-Seq dataset in appendix), which is attributable to both biological and technological factors. As a consequence, simple summary statistics measures such as the total number of reads within the exon boundary will be adversely affected by these complications. To overcome these problems, we propose to model the sequencing depth at each base accounting for the spatial correlation. These models explicitly take into consideration the variation and the correlation; therefore they are capable of facilitating proper normalization on sequencing depth at the base-level.

Let Y_{jk} represent the sequencing depth at the j th base of the specified exon in the k th sample in a study. Here $j = 1, \ldots n$, and n is the length of the exon; $k = 1, \ldots, m$, and m is the total number of replicated samples. Our goal is to build a probability model, analogous to the one proposed in Li and Wong (2001) (111) for modeling the probe-level microarray data, to capture the observed variation in Y_{jk}. There are many factors, both biological and technological, that affect Y_{jk}. In order to keep the model simple and to avoid overfitting, we propose to use spatial Poisson regression models.

At least five different categories of factors contribute to the variation of the reads Y_{jk}. The most important category, also the one we are most interested in estimating, consists of the biological factors that represent the expression level of this exon in the cell type under study. Examples of cell types are treated or untreated, disease or normal.

Suppose there are in total p such biological factors in the k th sample, denoted by $X_{k1}, X_{k2}, \ldots, X_{kp}$. Let $\boldsymbol{X_k} = (X_{k1}, X_{k2}, \ldots, X_{kp})$.

The second category represents factors that are specific to the current sample. They arose from the sample preparation step and sequencing process, hence are referred to as technical factors. An example is the lane effect. Compared to other factors, studies have shown that variation introduced by technical factors is typically not significant (107). Suppose there are in total q such technical factors in the k th sample, denoted by $T_{k1}, T_{k2}, \ldots, T_{kq}$. Let $\boldsymbol{T_k} = (T_{k1}, T_{k2}, \ldots, T_{kq})$.

The third category includes systematic biases introduced by intrinsic local genomic features, such as GC content that affect sequencing depth. This type of bias is platform-specific. As an example, Dohm et al. (2008) (112) showed that the sequencing depth is positively correlated with GC content using the Illumina/Solexa platform. We use a generic term r_j to represent biases that belong to this category. In the preliminary analysis of a real RNA-Seq dataset (appendix), we observed a quadratic association pattern between sequencing depth and GC content. Therefore we add both first and second order terms of GC content in the model.

The factors in the first three categories are considered fixed. The last two categories contain random effects. The fourth category consists of random effects $V_{1k}, V_{2k}, \ldots V_{nk} (1 \leq k \leq m)$ at base 1 through n in the k th sample, respectively. As in spatial statistics, V_{jk}'s are assumed to be independent and identically distributed as $N(0, \sigma_v^2)$. These effects are used to account for unstructured variability, which may be attributed to some latent factors of over-dispersion.

The last category includes $U_{1k}, U_{2k}, \ldots, U_{nk} (1 \leq k \leq m)$ representing the spatial correlation between the base counts. Like in spatial statistics, there are many ways to specify this spatial correlation structure. In this study, we will primarily investigate two configurations, the joint structure and the conditional or ICAR structure.

One way to specify the spatial correlation structure of U_{jk}'s in the k th sample is to assume that they follow the same multivariate Gaussian distribution $N(0, \sigma_u^2 \Sigma)$, where $\Sigma = (\sigma_{ij})$ is the correlation matrix of U_{jk}'s and $\sigma_{ij} = Corr(U_{ik}, U_{jk})$. It is usually not possible to estimate a general covariance matrix Σ. Often, σ_{ij} is assumed to be a parametric function of distance d_{ij} between the i th base and the j th base. Clearly, the parametric function cannot be arbitrary and is required to result in a positive definite matrix Σ. One popular choice is $\sigma_{ij} = \rho^{d_{ij}}$ where ρ is a parameter determining the extent of correlation. This structure is referred to as the joint structure.

Another approach is to impose the correlation structure locally using a Gaussian Markov random field. This idea was originally proposed by Besag (1974) (113) and others, and was used by both Clayton and Kaldor (1987) (114) and Besag et al. (1991) (115). For a fixed base j, first we need to define its neighborhood ∂j. A simple approach is to define ∂j as the collection of every base i which is adjacent to base j. Other definitions are possible, for example, Cressie and Chan (1989) (116) defines neighborhoods in terms of the distance d_{ij}. Second, we need to define a weight matrix $W = (w_{ij})$ as follows. For $1 \leq i, j \leq n, w_{ii} = 0$; $w_{ij} = 1$ if $i \in \partial j$; and $w_{ij} = 0$ otherwise. Let $U_{(-j)k}$ denote the collection of U_{ik} 's with $i \neq j$. The conditional distribution of U_{jk} given $U_{(-j)k}$ is assumed to be:

$$U_{jk}|U_{(-j)k} \sim Normal\left(\frac{\sum_{i\in\partial j} w_{ij}U_{ik}}{\sum_i w_{ij}}, \frac{\sigma_u^2}{\sum_i w_{ij}}\right) \quad (4.1)$$

The correlation is referred to as the Gaussian Markov structure or the intrinsic conditional autoregressive (ICAR) structure.

The discussion about the pros and cons of the joint and ICAR structures can be found in Best el al. (2005) (117) and Wakefield (2007) (118). Other more sophisticated models are also available in the literature. Two examples are the mixture model proposed by Green and Richardson (2002) (119) and the spatial partition model proposed by Knorr-Held and Raber (2000) (120). Best et al. (2005) (117) presents a thorough simulation study for comparing the performances of different models, and concludes that both the joint and ICAR structures have good properties for disease mapping. Covariates and deterministic spatial trends can be incorporated into the two structures in a straightforward fashion.

In the literature on disease mapping, MCMC methods are the predominant methods used for model fitting and inference, following Besag et al. (1991) (115) and Diggle et al. (1998) (121). In the literature on imaging processing, however, maximum likelihood methods are also used; see Zhu et al. (2009) (122) for example. The preference for Bayesian computational methods is due to the fact that the spatial random effects U_{jk}'s and V_{jk}'s are not directly observable; and it takes high-dimensional integration to integrate them out, which can be computationally challenging. In the study, we will follow the tradition of using Bayesian computational methods when we apply the spatial models to the RNA-Seq data analysis.

We propose the following spatial Poisson regression model for RNA-Seq data of a fixed exon:

$$\begin{aligned} & Y_{jk} | \theta_k, U_{jk}, V_{jk}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta} \\ & \sim Poisson\left(n\theta_k \exp\left\{\boldsymbol{X_k\gamma} + \boldsymbol{T_k\eta} + \alpha r_j + \boldsymbol{\beta} r_j^2 + U_{jk} + V_{jk}\right\}\right) \end{aligned} \tag{4.2}$$

Where θ_k, α, $\boldsymbol{\beta}$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots \gamma_p)$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots \eta_q)$ are the model parameters. In addition, the specifications of the random effects also include more parameters, depending on which configuration is adopted.

When the joint structure is assumed for U_{jk}'s, then the model specified by (4.2) is referred to as the joint spatial Poisson regression model; when the ICAR structure is assumed for U_{jk}'s, then the model specified by (4.2) is referred to as the ICAR spatial Poisson regression model. Note that the weight matrix W has to be further specified when we use ICAR structure for U_{jk}'s. In this study, we define the neighborhood ∂j as the adjacent bases. If we not only include the bases directly adjacent to j but also include bases that are within d bases of j, assigning equal weights to bases in ∂j may be not appropriate, therefore other weighting schemes need to be considered.

The expression level of the exon in the k th sample under biological factor $\boldsymbol{X_k}$ can be defined as $n\theta_k \exp\{\boldsymbol{X_k\gamma}\}$, which is adjusted by GC content bias term r_j, the technical factor $\boldsymbol{T_k}$, the location-specific random effect V_{jk} and the spatial random effect U_{jk}. From the model the term $n\theta_k \exp\{\boldsymbol{X_k\gamma}\}$ can be used as the normalized expression level for comparison between different exons and between different cell lines.

4.2.2 Model implementation

We adopt a Bayesian approach and use MCMC methods to carry out estimation and inference. We start by assigning appropriate priors for model parameters. The

marginal distribution of random effects U_{jk} and V_{jk} are $N(0, \sigma_u^2)$ and $N(0, \sigma_v^2)$, respectively. If we specify inverse Gamma priors for σ_u^2 and σ_v^2, then the implied prior for $\sigma_u^2 + \sigma_v^2$ is not inverse Gamma so that we cannot easily control the total variance. Following the notations in Wakefield (2007) (118), we write the total precision as: $\tau = (\sigma_u^2 + \sigma_v^2)^{-1}$, and specify a Gamma prior $\tau \sim \Gamma(a, b)$. Then let $p = \sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ represent the proportion of the total variation that is attributable to the spatial component, and assign a Beta prior, $Beta(c, d)$, to p, and transform from (τ, p) to (σ_u^2, σ_v^2) via:

$$\sigma_u^2 = \frac{p}{\tau} = p(\sigma_u^2 + \sigma_v^2), \qquad \sigma_v^2 = \frac{1-p}{\tau} = (1-p)(\sigma_u^2 + \sigma_v^2) \tag{4.3}$$

If the joint structure is assumed for U_{jk}'s, there is another parameter ρ determining the extent of global correlation. We assign another Beta prior, $Beta(e, f)$, to ρ. We assign non-informative prior $I(0, +\infty)$ for θ_k, and $N(\mu_\alpha, \sigma_\alpha^2)$, $N(\mu_\beta, \sigma_\beta^2)$ for α and β.

In the default setting, we set uniform priors for p and ρ with $c = d = e = f = 1$. In addition, we use weakly informative priors for α and β with $\mu_\alpha = \mu_\beta = 0$, $\sigma_\alpha^2 = \sigma_\beta^2 = 10{,}000$. It is tricky to assign a prior for the total precision $\tau = (\sigma_u^2 + \sigma_v^2)^{-1}$. We use an empirical Bayesian method: use moment estimate $\hat{\tau}$ to assign an informative prior $\Gamma(a, b)$ for τ. Specifically, we consider a simplified joint spatial Poisson regression model with no biological factors, technological factors or systematic biases.

$$Y_{jk} | \theta_k, U_{jk}, V_{jk} \sim Poisson\left(n\theta_k \exp\{U_{jk} + V_{jk}\}\right) \tag{4.4}$$

The marginal distribution of random effects U_{jk} and V_{jk} are $N(0, \sigma_u^2)$ and $N(0, \sigma_v^2)$, respectively. U_{jk} and V_{jk} are independent. It is straightforward to calculate the first moment and the second moment of Y_{jk}:

$$EY_{jk} = E\{E(Y_{jk}|U_{jk}, V_{jk})\} = n\theta_k Ee^{U_{jk}} Ee^{V_{jk}} = n\theta_k e^{1/2\tau} \quad (4.5)$$

$$\begin{aligned} EY_{jk}^2 &= E\{E(Y_{jk}^2|U_{jk}, V_{jk})\} \\ &= n\theta_k Ee^{U_{jk}} Ee^{V_{jk}} + n^2\theta_k^2 Ee^{2U_{jk}} Ee^{2V_{jk}} \\ &= n\theta_k e^{1/2\tau} + n^2\theta_k^2 e^{2/\tau} \end{aligned} \quad (4.6)$$

Then we can get the moment estimate $\hat{\tau}_k$ for total precision τ:

$$\hat{\tau}_k = 1/log\left(\frac{\widehat{EY_{jk}^2} - \widehat{EY_{jk}}}{\left(\widehat{EY_{jk}}\right)^2}\right) \quad (4.7)$$

where

$$\widehat{EY_{jk}} = \frac{1}{n}\sum_{j=1}^{n} Y_{jk}, \widehat{EY_{jk}^2} = \frac{1}{n}\sum_{j=1}^{n} Y_{jk}^2 \quad (4.8)$$

To avoid taking the logarithm of a negative number, we only work on exons with $\widehat{EY_{jk}^2} > \widehat{EY_{jk}}$ (see over-dispersion filter in real data analysis section). We fix the coefficient of variation (CV: ratio of the standard deviation to the mean) to be 1 in the informative prior $\Gamma(a, b)$ for τ by setting $a = 1$ and $b = max_{1\le k\le m}\,\hat{\tau}_k$.

It is straightforward to derive the joint posterior distribution (formulas in appendix), and we use Gibbs sampler to iteratively sample parameters from the conditional posterior distribution. Since we use conjugate priors, the conditional posterior distributions for θ_k and τ are Gamma distributions, which are easy to sample. The conditional posterior distributions for fixed effects α and β, and random effects U_{jk}'s and V_{jk}'s are complicated and not in closed forms. Fortunately, they are all log-concave functions, and we can use adaptive rejection sample (ARS) (123) to draw samples. For the other parameters, p and ρ, their conditional posterior distributions are not log-concave, but very nearly log-concave. We use adaptive rejection Metropolis-Hastings

sampling (ARMS) (124) to draw samples from these complicated distributions. One important issue in using ARS and ARMS is to assign appropriate ranges for the parameters. In ARS, we use interval $[-100, 100]$ for fixed effects α and β, and interval $[-5, 5]$ for random effects U_{jk}'s and V_{jk}'s, respectively. In ARMS, we use the natural boundary $[0, 1]$ for the parameters, p and ρ.

In addition, to make the intercept term θ_k's identifiable, we add two constraints on the random effects in each iteration of the Gibbs sampler:

$$\sum_{j=1}^{n} U_{jk} = 0, \sum_{j=1}^{n} V_{jk} = 0, 1 \le k \le m \tag{4.9}$$

To be specific, at the end of each iteration, we replaced U_{jk} and V_{jk} by $U_{jk} - \bar{U}_{.k}$ and $V_{jk} - \bar{V}_{.k}$.

4.3 Simulation study

In this study, we propose the spatial Poisson regression model for exon level RNA-Seq data. This model is able to capture spatial correlation and non-spatial variation, and we believe the expression index estimated from such model will reflect the underlying expression levels more accurately. The reads per kilobase of exon per million mapped reads (RPKM) proposed by Mortazavi et al. (2008) (104) is one of the most popular existing methods to measure exon level gene expression. RPKM takes into consideration exon length and the total read number in the measurement, and it can be calculated as the normalized sample mean. Another similar approach is to use the normalized sample median (the median number of reads within an exon per kilobase of exon model per million mapped reads), which is robust to sporadic outliers. Next we conducted a series of simulation studies to compare the exon level gene expression

measurement estimated by the spatial Poisson regression models, the normalized sample mean method and the normalized sample median method.

To make the simulation study simple and still without loss of generality, we did not consider the technical factors. Further we only incorporated one biological factor: cases vs. controls. We assumed that the cases and the controls are independent and have distinct parameters; therefore we can simulate them independently.

4.3.1 No systematic bias

We started from a simple case with no systematic bias. Let Y_{ijk} represent the sequencing depth at the j th base of a specified exon where $j = 1, \ldots, n$, and $n = 100$ is the length of the exon; $k = 1, \ldots, m$, and $m = 7$ is the number of duplicated samples. Here i is an index for the experimental conditions: $i = 1$ indicates the case sample, and $i = 0$ indicates the control sample. We simulated the seven cases and the seven controls from a joint spatial Poisson regression model separately:

$$\begin{gathered} Y_{ijk} | \theta_{ik}, U_{ijk}, V_{ijk} \sim Poisson\left(n\theta_{ik} \exp\{U_{ijk} + V_{ijk}\}\right) \\ U_{i1k}, U_{i2k}, \ldots, U_{ink} \sim N(0, \Sigma_{ui}), Var(U_{ijk}) = \sigma_{ui}^2, Cor\left(U_{ij_1k}, U_{ij_2k}\right) = \rho^{|j_1 - j_2|} \\ V_{ijk} \; iid \sim N(0, \sigma_{vi}^2) \end{gathered} \quad (4.10)$$

We need to specify parameters: θ_{0k}, θ_{1k}, σ_{u0}^2, σ_{u1}^2, σ_{v0}^2, σ_{v1}^2 and ρ in the simulation. First assuming that expression levels in the cases and the controls were the same ($\theta_{1k} = \theta_{0k} = 1$), and then the true fold change was $\theta_{1k}/\theta_{0k} = 1$. ρ was set to 0.9 since the real RNA-Seq data exhibited strong spatial correlations (see preliminary analysis of a real RNA-Seq dataset in appendix). To mimic distinct variability observed in the real RNA-Seq data (Figure A4.1 in appendix), we designed five combinations of the variance of two random effects, σ_{u0}^2, σ_{u1}^2, σ_{v0}^2 and σ_{v1}^2 (Table 4.1).

Table 4.1 Five combinations of variance of two random effects.

Settings	Cases		Controls	
	Spatial variance	Non-spatial variance	Spatial variance	Non-spatial variance
A	0.04	0.01	0.04	0.01
B	0.05	0.02	0.04	0.01
C	0.04	0.01	0.05	0.02
D	0.05	0.01	0.04	0.02
E	0.04	0.02	0.05	0.01

In setting A, the total variances of two random effects are the same in the cases and the controls, and the ratios between spatial and non-spatial variances are the same too. In setting B, the total variance of two random effects in the cases is larger than that in the controls. In setting C, the total variance of two random effects in the cases is smaller than that in the controls. In setting D, the total variances of two random effects are the same in the cases and the controls, but the ratio between spatial and non-spatial variances in the cases is larger than that in the controls. In setting E, the total variances of two random effects are the same in the cases and the controls, but the ratio between spatial and non-spatial variances in the cases is smaller than that in the controls.

We estimated expression indexes for the cases and the controls, and the corresponding fold change (ratio) using four methods: the joint spatial Poisson model, the ICAR spatial Poisson model, the normalized sample mean method and the normalized sample median method. For two spatial Poisson models we proposed, we used 10,000 iterations in each Gibbs sample. The first 9,000 samplers were dropped as the burn-in stage, and then every 10^{th} sample in the last 1,000 samples were used to calculate the posterior mean. We ran five parallel chains and used the one with highest posterior mode.

We repeated 100 simulations for each of five cases, and recorded the mean estimate and the mean square error (MSE) across 100 simulations (Table 4.2).

Table 4.2 Simulation results when there is no systemic bias. Best methods (smallest MSE) in each setting are listed in bold font.

Settings	Methods	Case		Control		Fold change	
		Estimate	MSE	Estimate	MSE	Estimate	MSE
A	Joint	1.000354	**0.000014**	0.999733	**0.000019**	1.000642	**0.000036**
	ICAR	1.000749	0.000015	0.999944	**0.000019**	1.000826	0.000037
	mean	1.022173	0.000514	1.021895	0.000508	1.000302	0.000052
	median	0.999950	0.000070	0.998843	0.000066	1.001178	0.000147
B	Joint	1.000660	**0.000015**	1.000020	**0.000016**	1.000656	**0.000032**
	ICAR	1.000379	**0.000015**	1.000253	**0.000016**	1.000143	**0.000032**
	mean	1.031257	0.001004	1.021648	0.000493	1.009431	0.000142
	median	0.999607	0.000080	1.000014	0.000081	0.999686	0.000187
C	Joint	1.000329	**0.000015**	1.000202	**0.000013**	1.000139	**0.000026**
	ICAR	1.000540	**0.000015**	0.999944	**0.000013**	1.000609	**0.000026**
	mean	1.022173	0.000513	1.030123	0.000929	0.992300	0.000094
	median	0.999771	0.000072	0.999457	0.000080	1.000398	0.000158
D	Joint	1.000503	**0.000015**	1.000829	**0.000014**	0.999688	**0.000029**
	ICAR	1.000698	0.000016	1.000083	**0.000014**	1.000630	0.000031
	mean	1.026408	0.000721	1.027008	0.000754	0.999441	0.000047
	median	1.000614	0.000071	1.000464	0.000085	1.000255	0.000195
E	Joint	1.000507	**0.000016**	1.000225	0.000015	1.000296	**0.000030**
	ICAR	1.000003	0.000017	1.000498	**0.000014**	0.999519	0.000032
	mean	1.027096	0.000762	1.026273	0.000717	1.000830	0.000056
	median	0.999164	0.000069	1.000129	0.000095	0.999139	0.000182

As we expected from a simple calculation (appendix), the normalized sample mean method overestimated the expression level of the cases and the controls (1.0216 ~ 1.0313) in all five settings. In setting A, D and E, the normalized sample mean method

provided unbiased estimates of the fold change (1.0003, 0.9994, 1.0008). However, in setting B, the normalized sample mean method overestimated the fold change (1.0094), and in setting C, the normalized sample mean method underestimated the fold change (0.9923). Compared to the normalized sample mean method, the normalized sample median method provided unbiased and efficient mean estimates (smaller MSE) in all five settings. However, the fold changes estimated in the normalized sample median method were unbiased but inefficient (larger MSE). The performance of the joint spatial Poisson model and the ICAR spatial Poisson model were similar. The two spatial Poisson models provided more efficient estimates (smaller MSE) of the expression level and the fold change compared to the normalized sample mean method and the normalized sample median method.

4.3.2 Adding GC content as systematic bias

We observed a quadratic pattern between GC content and sequencing depth from an exploratory analysis of a real RNA-Seq dataset (appendix). It is challenging to measure the true underlying gene expression level with such systematic bias. To test the impact of GC content, we extended the previous simulation study be adding GC content as covariates in the mean structure. Here we compared four methods: the joint spatial Poisson model with GC content as covariates (Joint_GC), the ICAR spatial Poisson model with GC content as covariates (ICAR_GC), the normalized sample mean method (mean) and the normalized sample median method (median).

We first simulated seven cases and seven controls from the joint spatial Poisson regression model separately:

$$Y_{ijk}|\theta_{ik}, U_{ijk}, V_{ijk} \tag{4.11}$$

$$\sim Poisson\left(n\theta_{ik}\exp\left\{\alpha_i*(r_j-\bar{r})+\beta_i*(r_j-\bar{r})^2+U_{ijk}+V_{ijk}\right\}\right)$$
$$U_{i1k},U_{i2k},\ldots,U_{ink}\sim N(0,\Sigma_{ui}),Var(U_{ijk})=\sigma_{ui}^2,Cor(U_{ij_1k},U_{ij_2k})=\rho^{|j_1-j_2|}$$
$$V_{ijk}\ iid\sim N(0,\sigma_{vi}^2)$$

In this model, r_j is the GC content at the j th base, and $\bar{r}$ is the average GC content in this specified exon. We simulated r_j from a normal distribution with mean 0.5373 and standard error 0.1197. These two values are calculated from the real data. (see preliminary analysis of a real RNA-Seq data in appendix). Note that the GC content is the same at the j th base across all duplicated samples. Further we need to specify parameters: θ_{0k}, θ_{1k}, σ_{u0}^2, σ_{u1}^2, σ_{v0}^2, σ_{v1}^2, ρ, α_0, α_1, β_0 and β_1. The same simulation setting was adopted here: $\theta_{0k}=\theta_{1k}=1$, $\rho=0.9$ and the true fold change is $\theta_{1k}/\theta_{0k}=1$. The variance of two random effects, σ_{u0}^2, σ_{u1}^2, σ_{v0}^2, σ_{v1}^2 were also from Table 4.1. To mimic the real RNA-Seq data (see preliminary analysis of a real RNA-Seq dataset in appendix), we set $\alpha_0=5.0531$, $\alpha_1=5.5556$, $\beta_0=-16.4003$ and $\beta_1=-15.8962$. Here we assumed that the true expression level $n\theta_{ik}$ was the sequencing depth at some base with GC content $\bar{r}$. In the simulation, we adopted a quadratic function form between GC content and sequencing depth, therefore the sequencing depth at most of the bases will be smaller than $n\theta_{ik}$, and both the normalized sample mean method and the normalized sample median method underestimated the true expression level. Such bias was larger in controls than in cases since the coefficient (β_0) of second order term $(r_j-\bar{r})^2$ in the controls was smaller than those (β_1) in the cases. As a consequence, both the normalized sample mean method and the normalized sample median method overestimated the true fold change. We used the same approach as the previous simulation study to record the simulation results and compared the performance among four different methods. The

values in Table 4.3 are the mean estimate and the mean square error (MSE) across 100 simulations.

Table 4.3 Simulation results when there is systemic bias. Best methods (smallest MSE) in each setting are listed in bold font.

Settings		Case		Control		Fold change	
		Estimate	MSE	Estimate	MSE	Estimate	MSE
A	Joint_GC	0.9993	**0.0002**	1.0019	**0.0002**	0.9976	**0.0006**
	ICAR_GC	0.9944	0.0004	1.0006	0.0004	0.9942	0.0007
	Mean	0.9838	0.0007	0.9528	0.0027	1.0328	0.0015
	Median	0.9639	0.0037	0.9550	0.0038	1.0098	0.0018
B	Joint_GC	1.0054	**0.0003**	1.0036	**0.0003**	1.0020	**0.0005**
	ICAR_GC	1.0005	0.0006	0.9955	**0.0003**	1.0053	0.0008
	Mean	0.9951	0.0005	0.9578	0.0023	1.0392	0.0021
	Median	0.9646	0.0030	0.9620	0.0030	1.0034	0.0017
C	Joint_GC	1.0035	**0.0003**	0.9999	**0.0004**	1.0040	**0.0007**
	ICAR_GC	1.0017	0.0004	0.9987	0.0008	1.0036	0.0010
	Mean	0.9880	0.0006	0.9648	0.0017	1.0242	0.0010
	Median	0.9760	0.0021	0.9430	0.0046	1.0357	0.0029
D	Joint_GC	1.0020	**0.0004**	1.0033	**0.0004**	0.9990	**0.0007**
	ICAR_GC	1.0007	0.0006	0.9999	**0.0004**	1.0011	0.0009
	Mean	0.9908	0.0005	0.9567	0.0022	1.0358	0.0017
	Median	0.9657	0.0031	0.9516	0.0037	1.0156	0.0026
E	Joint_GC	1.0022	**0.0003**	1.0015	**0.0003**	1.0009	**0.0006**
	ICAR_GC	0.9988	0.0005	1.0024	0.0006	0.9971	0.0012
	Mean	0.9915	0.0005	0.9603	0.0019	1.0326	0.0014
	Median	0.9620	0.0032	0.9530	0.0035	1.0101	0.0019

We observed that both the normalized sample mean method and the normalized sample median method underestimated the expression level of the cases and the controls in all five settings. In addition, both the normalized sample mean method and the normalized sample median method overestimated the fold change in all five settings.

With GC content as covariates, in all five settings both the joint spatial Poisson model and the ICAR spatial Poisson model provided more efficient estimate (smaller MSE) of expression level of the cases and controls, and the corresponding fold change. Further, the joint spatial Poisson model was better than the ICAR spatial Poisson model in terms of MSE. That was what we expected since the datasets were simulated from the joint spatial Poisson model.

In the joint spatial Poisson model with GC content as covariates (Joint_GC) and the ICAR spatial Poisson model with GC content as covariates (ICAR_GC), we can obtain the estimate of GC content effects. To compare these two methods, we also recorded the bias and the mean square error (MSE) of α_0, α_1, β_0 and β_1 across 100 simulations in Table 4.4.

The joint spatial Poisson model with GC content as covariates (Joint_GC) provided more efficient estimates (smaller MSE) of GC content effects than the ICAR spatial Poisson model with GC content as covariates (ICAR_GC).

Table 4.4 Bias and MSE of GC content effects

Settings	Methods	α_1		β_1		α_0		β_0	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
A	Joint_GC	-0.0389	0.0708	0.0808	1.2813	-0.0375	0.0565	-0.0526	1.0827
	ICAR_GC	-0.5335	0.5996	0.4426	2.3226	-0.4451	0.4779	0.0088	2.3588
B	Joint_GC	-0.0544	0.0701	-0.2492	1.6509	0.0073	0.0787	-0.2457	1.4625
	ICAR_GC	-0.5152	0.5852	0.0534	2.9662	-0.4057	0.4626	0.3396	1.7157
C	Joint_GC	0.0103	0.0767	-0.2067	1.3550	-0.0218	0.0868	0.0890	2.0289
	ICAR_GC	-0.3835	0.4905	-0.0727	1.8915	-0.5484	0.6385	0.1682	4.0081
D	Joint_GC	0.0281	0.1028	-0.1476	2.0763	-0.0371	0.0803	-0.2386	1.8750
	ICAR_GC	-0.4127	0.4854	-0.0727	3.3440	-0.4287	0.4698	0.0115	2.1239
E	Joint_GC	-0.0548	0.0874	-0.1105	1.7161	-0.0391	0.0773	-0.0785	1.4506
	ICAR_GC	-0.5020	0.5495	0.1330	2.8561	-0.4721	0.5326	-0.1578	3.5296

4.4 Real data analysis

In this study we analyzed the sequencing quality control project (SEQC) dataset (unpublished data from Dr Jun Li, Department of Human Genetics, University of Michigan), which consists of two samples from microarray quality control project (MAQC) (125): brain and universal human reference (UHR). They were prepared using the standard Illumina mRNA-Seq protocol and reagents and sequenced across seven lanes (across two single flowcells) to a depth of 35 bases. The data were generated using Illumina pipeline 1.1 and eland_rna which is optimized for alignment of transcriptomic reads. Only reads that uniquely mapped to the human reference genome, with up to two mismatches, were included in this study. Total numbers of uniquely mapped reads are listed in Table 4.5 (Lane 5 was used as the negative control).

Table 4.5 Total numbers of uniquely mapped reads in the SEQC dataset

Lane	Brain (million)	UHR (million)
Lane 1	5.4285	6.5531
Lane 2	6.3744	6.6063
Lane 3	6.4921	7.2336
Lane 4	6.6011	7.1439
Lane 6	6.6963	7.0093
Lane 7	6.5739	6.8177
Lane 8	6.1030	6.4989

Exon level annotations for 385,122 ENSEMBL exons were downloaded from ENSEMBL database version 55: http://jul2009.archive.ensembl.org/biomart/martview/ with genome assembly GRCh37. Using UCSC genome browser tool "liftover" to convert GRCh37 to NCBI36, total number of exons was reduced to 384,763, and then the base-

level sequencing depth files were generated for these 384,763 exons in seven brain samples and seven UHR samples.

Three exon filters were adopted in this study (Table A4.1 in appendix). The first one was the non-zero filter. First we removed all exons with zero sequencing depth since they contained no information. 170,877 exons left in the brain samples, 185,253 exons left in the UHR samples. The second one was the over-dispersion filter. The observed sequencing depth files exhibited substantial over-dispersion (Figures A4.1 in appendix): the variance of the number of reads in each exon is larger than the mean of the number of reads in each exon in the majority of exons with non-zero read coverage. We measured such over-dispersion in a specified exon by the test statistic proposed in Dean (1992) (126):

$$t_{ik} = \frac{\sum_{j=1}^{n}\left(\left(y_{ijk} - \bar{y}_{ik}\right)^2 - \bar{y}_{ik}\right)}{\sqrt{2n}\,\bar{y}_{ik}}, \qquad \bar{y}_{ik} = \frac{1}{n}\sum_{j=1}^{n} y_{ijk} \tag{4.12}$$

Here i is an index for the experimental conditions: $i = 1$ indicates the brain samples, and $i = 0$ indicates the UHR samples. $j = 1, \ldots n$ is an index for base, and n is the length of the specified exon. $k = 1, \ldots, m$ is an index for sample, and $m = 7$ is the number of duplicated samples. Under H_0 when y_{ijk} was generated from a Poisson distribution, t_{ik} followed the standard normal distribution. Using the over-dispersion filter, we selected a subset of over-dispersed exons ($t_{1k} \geq 1.96$ in all seven brain samples and $t_{0k} \geq 1.96$ in all seven UHR samples). After this filtering, 43,141 exons were left in the brain samples and 56,947 exons were left in the UHR samples. The third one was the "non-gap" filter. Ideally, we expected that the sequencing depth would be constant within each exon. However, we observed some gaps (zero read coverage) in the long exons (see

preliminary analysis of a real RNA-Seq dataset in appendix). Such gaps may be due to low overall sequencing depth, repeated regions, or inaccurate exon annotations. To avoid such complicating issues, we adopted a very stringent "non-gap" filter in this study by defining "non-gap" exons as those exons without gaps in either one of seven samples. 3,569 "non-gap" over-dispersed exons in brain samples and 5,241 "non-gap" over-dispersed exons in UHR samples were left after the three filters.

Three methods, the normalized sample mean method, the joint spatial Poisson model without GC content as covariates and the joint spatial Poisson model with GC content as covariates, were used to measure the exon level gene expression in 3,569 exons in the brain samples and 5,241 exons in the UHR samples. The three methods were referred to as mean, noGC and GC in the following tables (Table 4.6, 4.7, 4.8, 4.9 and 4.10), respectively. Due to the intensive computation, we used one chain in each Gibbs sample with 1,000 iterations. The first 900 samples were dropped as the burn-in stage, and then every 10^{th} sample in the last 100 samples were used to calculate the posterior mean.

We compared the RNA-Seq data with the exon array data. The exon array data (127) was download from the NCBI Gene Expression Omnibus (GEO) repository under the GEO records GSE13072. Affymetrix GeneChip Human Exon 1.0 ST arrays were used to measure the brain samples and the UHR samples. There were ten brain samples and ten UHR samples, respectively. The custom CDF (128) file was used to process the raw CEL files, and then we obtained the exon level gene expression data for 304,495 ENSEMBL exons. Merging the RNA-Seq data with the exon array data, 3,175 exons in the brain samples and 4,499 exons in the UHR samples were left.

We used the exon array data as the gold standard, and compared the three methods: mean, noGC and GC, by Spearman's rank correlation coefficients (Table 4.6). The overall performances were very similar. The normalized sample mean method was the best, while the joint spatial Poisson model with GC content as covariates was slightly better than the joint spatial Poisson model without GC content as covariates.

Table 4.6 Spearman's rank correlation coefficients for three methods

Sample	No. of exons	Mean	noGC	GC
Brain	3,175	0.4976	0.4915	0.4954
UHR	4,499	0.4579	0.4523	0.4550

Inspired by the ideas in Li et al (129), we looked into the reason why our spatial Poisson models failed to provide more accurate exon level gene expression estimates. We believe that in most of the exons we compared, the three estimates were very similar. Therefore there was little improvement in the overall Spearman's rank correlation coefficients. To see whether the joint spatial Poisson model can lead to improvement in those cases when it was different from the normalized sample mean method, we defined the distance d as the absolute log fold change between the normalized sample mean method and the joint spatial Poisson model with GC content as covariates:

$$d = \left| log\left(\frac{RPKM\_Mean}{Joint\_GC}\right) \right| \tag{4.13}$$

The distance d can quantify the difference between the estimates from the normalized sample mean method and from the joint spatial Poisson model with GC content as covariates. We further classified all exons into four groups according to d, and then compared the performance of the three methods within each group (Table 4.7).

Table 4.7 Spearman's rank correlation coefficients for three methods at difference d levels.

Sample	Distance	# of exons	Mean	noGC	GC	noGC_Mean[1]	GC_Mean[2]	GC_noGC[3]
Brain	d>0.04	34	0.3553	0.4874	0.4933	37.17%	38.76%	1.16%
	d>0.03	113	0.2397	0.2689	0.2745	12.17%	14.52%	2.09%
	d>0.02	455	0.4962	0.506	0.5099	1.99%	2.76%	0.76%
	d>0.01	1531	0.4904	0.4871	0.4932	-0.66%	0.54%	1.21%
UHR	d>0.04	50	0.5138	0.5408	0.5636	5.25%	9.69%	4.22%
	d>0.03	169	0.3188	0.3222	0.3419	1.07%	7.26%	6.13%
	d>0.02	692	0.4921	0.4918	0.5042	-0.03%	2.50%	2.53%
	d>0.01	2310	0.4956	0.4937	0.4957	-0.39%	0.03%	0.42%

[1]noGC_Mean: relative improvement between the joint spatial Poisson model without GC content as covariates and the normalized sample mean method.

[2]GC_Mean: relative improvement between the joint spatial Poisson model with GC content as covariates and the normalized sample mean method.

[3]GC_noGC: relative improvement between the joint spatial Poisson model with GC content as covariates and the joint spatial Poisson model without GC content as covariates.

From the Table 4.7, we found that the joint spatial Poisson model achieved high improvement when the estimates from the joint spatial Poisson model and the estimates from the normalized sample mean method were different. When the distance d was large than 0.04, using the joint spatial Poisson model with GC content as covariates provided 38.76% relative improvement in the brain samples and 9.69% relative improvement in the UHR samples, in term of the Spearman's rank correlation coefficients. The relative improvement became smaller when the difference d were smaller. In addition, we found that using GC content as covariates was consistently better than without GC content.

To further investigate the benefit of modeling GC content, we quantified the difference between the joint spatial Poisson model with GC content and the joint spatial Poisson model without GC content by the distance d_{GC}: the absolute log fold change between two joint spatial Poisson models:

$$d_{GC} = \left| log\left(\frac{Joint\_GC}{Joint\_noGC}\right)\right| \tag{4.14}$$

The distance d_{GC} quantified the difference between two joint spatial Poisson models. We further classified all exons into four groups according to their distance d_{GC}, and then compared the performance of two joint spatial Poisson models within each group (Table 4.8).

We found that using GC content as covariates was consistently better than without GC content, especially when the difference between two methods was large. When the distance d_GC was large than 0.03, using the joint spatial Poisson model with GC content as covariates provided 11.52% relative improvement in the brain samples and 3.69%

relative improvement in the UHR samples, in term of the Spearman's rank correlation coefficients.

Table 4.8: Spearman's rank correlation coefficients for two joint spatial Poisson models at difference d_GC levels.

Sample	Distance	# of exons	noGC	GC	GC_noGC[1]
Brain	$d_{GC} > 0.03$	88	0.2835	0.3161	11.52%
	$d_{GC} > 0.02$	301	0.4199	0.4284	2.02%
	$d_{GC} > 0.01$	1248	0.4476	0.4530	1.21%
UHR	$d_{GC} > 0.03$	151	0.3477	0.3605	3.69%
	$d_{GC} > 0.02$	505	0.3630	0.3669	1.06%
	$d_{GC} > 0.01$	1795	0.4035	0.4079	1.09%

[1]GC_noGC: relative improvement between the joint spatial Poisson model with GC content as covariates and the joint spatial Poisson model without GC content as covariates.

4.5 Discussion

The recent arrival of next generation sequencing technologies is rapidly changing how we design future genetics and genomics studies (130,131). Compared to Sanger sequencing, these "next generation" methods can produce orders of magnitude more data, and can do so efficiently, accurately, and at a fraction of the cost. These attractive features motivate scientists to apply these sequencing technologies to a variety of applications. One of the successful applications is the transcriptome analysis or RNA-Seq (104,107). Compared to microarray, RNA-Seq offers clear advantages such as better dynamic range and ability to discover novel transcript. However, just like in the early days of microarray, scientists face daunting challenges when deriving gene expression

levels from the counts of sequencing reads. Current methods enumerate the number of reads landed within each exons and use the normalized read counts to represent the gene expression levels for that exon. However, the sequencing depth across exons varies substantially. The variation also differs across exons which will affect the simple enumeration method that is currently being used.

In microarray data analysis, estimates from parametric models such as the model-based expression index (MBEI) from dChip (111) proved to be much more accurate in reflecting the underlying expression levels than summary statistics of raw intensity values. In this study, borrowing the idea from model-based methods for analyzing microarray data, we develop model-based methods to analyze base-level sequencing depth data from RNA-Seq for each exon using spatial Poisson regression models. First we introduce two random effects for spatial correlation and non-spatial variation. In addition, we observe a quadratic GC content effect on sequencing depth, and add both linear and quadratic terms of GC content in the mean structure of the spatial Poisson regression models. Both simulation studies and real data analyses demonstrate that the expression index estimated from spatial Poisson regression models reflect the underlying expression levels more accurately, compared to existing methods that enumerate reads within each exon.

In this study, we propose spatial Poisson regression models for the exon level RNA-Seq data. It is easy to extend these models to isoform level and gene level RNA-Seq data. Since exons are nested within an isoform, one intuitive approach is to introduce a hierarchical structure into the models and model the spatial correlation between different exons within the same isoform. When there are multiple isoforms in a gene, some exons may belong to a number of different isoforms, and the expression levels of

these isoforms may also be different from each other. We can use a mixture of spatial Poisson models for gene level RNA-Seq data, and each spatial Poisson model will provide an isoform specific expression index. Further, the proportion of each isoform can be used to detect novel alternative splicing events.

We adopt three exon filters: non-zero filter, over-dispersion filter and non-gap filter in the real data analysis. It is reasonable to remove those exons without read coverage since there is no information for gene expression. Next, the spatial Poisson regression models implicitly assume that the observed exon level read counts to be over-dispersed by modeling two random effects. Among all exons, a majority of them exhibit significant over-dispersion, and our spatial Poisson regression models are designed for those over-dispersed exons. For those exons without significant over-dispersion, the Poisson regression or other well tailored models could be better alternative solutions. In addition, we can remove the exon with gaps by using the non-gap filter. Lots of systemic biases will create gaps with exon and make real RNA-Seq data very noisy. Some gaps could be introns or intergenetic regions due to incorrect exon annotations. Further masked genomic regions or repeated regions can't be mapped by short reads. The lowly expressed exons can't be measured with limited sequencing depth. In this study, we use the stringent non-gap filter to avoid such potential systemic biases, but more sophisticated models, such as zero-inflated Poisson model (101) or generalized Poisson model (100), may fit well for read counts in exons with gaps.

Using exon array data as the gold standard for exon level gene expression may be not the best choice, since it is well-known that array-based technologies suffer from limited dynamic range and high background noise. In the MAQC project (125),

quantitative real-time PCR (qRT-PCR) (132) has been used as the gold standard to measure the fold change of gene expression between brain samples and universal human reference samples. Further study (132) showed that 3' digital gene expression (DGE) data can provide high quality gene expression measurement. Using qRT-PCR and DGE data as the gold standard could be a good future direction. However, we only have gene level qRT-PCR and DGE data, the exon level qRT-PCR or DGE data is not available. Furthermore, it is challenging to directly compare them with RNA-Seq data due to alternative splicing.

There are still some limitations in the current work. First, the number of iterations used in the simulation study and real data analysis may not be large enough, and may result in biased estimates. Our spatial-based method is computational intensive compared to existing methods for analyzing RNA-Seq data. But since the same procedure is applied to each exon, the inference procedure can be run in parallel. We anticipate modern cluster computers will mitigate the computation burden we faced, especially with the massively parallel graphical processing unit (GPU) technology. We also believe that the accuracy in statistical inference outweighs computation cost. In addition, it is important to test the goodness of fit of our spatial-based method. Gelfand, Dey and Chang proposed a model determination approach using predictive distributions (133), and we will include this in the future research.

4.6 Appendix

4.6.1 Formulas used in the MCMC algorithm

4.6.1.1 Joint spatial Poisson regression model

Assume we have RNA-Seq data for a specific exon with length n. Let Y_{jk} represent the sequencing depth (number of mapped reads) at the j the base of a specified exon in the k th sample in a study. Here $j = 1, \dots, n$, and n is the length of the exon; $k = 1, \dots m$, and m is the number of replicated samples. There are many factors, both biological and technological, that affect Y_{jk} including: (1) biological factors, for example, GC content, labeled by r_j. (2) A random effect term to account for structure or spatial variability, denoted by U_{jk}. (3) A random effect term to account for unstructured, non-spatial variability, denoted by V_{jk}. We propose the following spatial Poisson regression model for RNA-Seq data of a fixed exon:

$$Y_{jk}|\theta_k, U_{jk}, V_{jk}, \alpha, \beta \sim Poisson\left(n\theta_k exp\left\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\right\}\right)$$

For the k th sample, we assume U_{jk} follow a multivariate normal distribution with mean 0 and variance matrix M, such that $Var(U_{jk}) = \sigma_u^2$ and $Corr(U_{j_1k}, U_{j_2k}) = \rho^{|j_1 - j_2|}$. The determinant and inverse of variance matrix M are:

$$|M| = \sigma_u^{2n}(1-\rho^2)^{n-1}, M^{-1} = \sigma_u^{-2}(1-\rho^2)^{-1}R$$

Where $R = \{r_{ij}\}$ is a 3-banded matrix: $r_{11} = r_{nn} = 1$; $r_{ii} = 1 + \rho^2, 2 \le i \le n-1$; $r_{ij} = -\rho, |i-j| = 1$; $r_{ij} = 0, |i-j| \ge 2$. In addition, we assume V_{jk} are independent and follow $N(0, \sigma_v^2)$. We adopt a Bayesian approach, and assign conjugate prior to some parameters. Define:

$$\tau = (\sigma_u^2 + \sigma_v^2)^{-1} \sim \Gamma(a, b), p(\tau) \propto \tau^{a-1} e^{-\frac{\tau}{b}}$$

$$p = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} \sim Beta(c, d), p(p) \propto p^{c-1}(1-p)^{d-1}$$

$$\rho \sim Beta(e, f), p(\rho) \propto \rho^{e-1}(1-\rho)^{f-1}$$

$$\alpha \sim N(\mu_\alpha, \sigma_\alpha^2), \beta \sim N(\mu_\beta, \sigma_\beta^2)$$

From the spatial Poisson regression model, we have:

$$P(\vec{Y} = \vec{y} | n, \vec{\theta}, r, \vec{U}, \vec{V}, \alpha, \beta) = \prod_{j=1}^{n} \prod_{k=1}^{m} P(Y_{jk} = y_{jk} | n, \theta_k, r_j, U_{jk}, V_{jk}, \alpha, \beta)$$

$$\propto \prod_{j=1}^{n} \prod_{k=1}^{m} \exp\left(-n\theta_k \exp\{\alpha r_j + \beta \mathrm{r}_\mathrm{j}^2 + U_{jk} + V_{jk}\}\right) \left(n\theta_k \exp\{\alpha r_j + \beta \mathrm{r}_\mathrm{j}^2 + U_{jk} + V_{jk}\}\right)^{y_{jk}}$$

$$\propto \exp\left(-n \sum_{j=1}^{n} \sum_{k=1}^{m} \theta_k \exp\{\alpha r_j + \beta \mathrm{r}_\mathrm{j}^2 + U_{jk} + V_{jk}\}\right)$$

$$\times \prod_{k=1}^{m} \theta_k^{\sum_{j=1}^{n} y_{jk}} exp\left(\sum_{j=1}^{n} \sum_{k=1}^{m} y_{jk} \exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\}\right)$$

For the k th sample,

$$P(\vec{U}_{\cdot k} | \sigma_u^2, \rho) \propto |M|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \vec{U}_{\cdot k}{}' M^{-1} \vec{U}_{\cdot k}\right\}$$

$$= (\sigma_u^2)^{-\frac{n}{2}} (1-\rho^2)^{-\frac{n-1}{2}} \exp\left\{-\frac{\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k}\right)^2 + (1-\rho^2) U_{n,k}^2}{2\sigma_u^2 (1-\rho^2)}\right\}$$

$$= p^{-\frac{n}{2}} \tau^{\frac{n}{2}} (1-\rho^2)^{-\frac{n-1}{2}} \exp\left\{-\frac{\tau \sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k}\right)^2 + (1-\rho^2) U_{n,k}^2}{2p(1-\rho^2)}\right\}$$

$$P(\vec{V}_{\cdot k} | \sigma_v^2) \propto (\sigma_v^2)^{-\frac{n}{2}} exp\left\{-\frac{\sum_{j=1}^{n} V_{jk}^2}{2\sigma_v^2}\right\} = (1-p)^{-\frac{n}{2}} \tau^{\frac{n}{2}} exp\left\{-\frac{\tau \sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

Then combine information in all m samples,

$$P(\vec{U} | \sigma_u^2, \rho) \propto \prod_{k=1}^{m} P(\vec{U}_{\cdot k} | \sigma_u^2, \rho)$$

$$= p^{-\frac{mn}{2}} \tau^{\frac{n}{2}} (1-\rho^2)^{-\frac{m(n-1)}{2}} \exp\left\{ -\frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} \right\}$$

$$P\left(\vec{V} \middle| \sigma_v^2\right) = \prod_{k=1}^{m} P\left(\vec{V}_{\cdot k} \middle| \sigma_v^2\right) \propto (1-p)^{-\frac{mn}{2}} \tau^{\frac{mn}{2}} exp\left\{ -\frac{\tau \sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}$$

The joint posterior probability is:

$$P\left(\vec{\theta}, \vec{U}, \vec{V}, \tau, p, \rho, \alpha, \beta \middle| \vec{Y}\right)$$

$$\propto P\left(\vec{Y} \middle| \vec{\theta}, \vec{U}, \vec{V}, \alpha, \beta\right) P\left(\vec{U} \middle| \tau, p, \rho\right) P\left(\vec{V} \middle| \tau, p\right) P(\tau) P(p) P(\rho) P\left(\vec{\theta}\right) p(\alpha) p(\beta)$$

$$= exp\left(-n \sum_{j=1}^{n} \sum_{k=1}^{m} \theta_k \, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} \right)$$

$$\times \prod_{k=1}^{m} \theta_k^{\sum_{j=1}^{n} y_{jk}} \, exp\left(\sum_{j=1}^{n} \sum_{k=1}^{m} y_{jk} (\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}) \right)$$

$$\times p^{-\frac{mn}{2}} \tau^{\frac{mn}{2}} (1-\rho^2)^{-\frac{m(n-1)}{2}} exp\left\{ -\frac{\tau \sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}$$

$$\times exp\left\{ -\frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} \right\} (1-p)^{-\frac{mn}{2}} \tau^{\frac{mn}{2}}$$

$$\times \tau^{a-1} e^{-\frac{\tau}{b}} p^{c-1} (1-p)^{d-1} \rho^{e-1} (1-\rho)^{f-1} exp\left\{ -\frac{(\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{(\beta - \mu_\beta)^2}{2\sigma_\beta^2} \right\}$$

After simplification, we can get:

$$P\left(\vec{\theta}, \vec{U}, \vec{V}, \tau, p, \rho, \alpha, \beta \middle| \vec{Y}\right)$$

$$\propto P\left(\vec{Y} \middle| \vec{\theta}, \vec{U}, \vec{V}, \alpha, \beta\right) P\left(\vec{U} \middle| \tau, p, \rho\right) P\left(\vec{V} \middle| \tau, p\right) P(\tau) P(p) P(\rho) P\left(\vec{\theta}\right) p(\alpha) p(\beta)$$

$$= exp\left(-n \sum_{j=1}^{n} \sum_{k=1}^{m} \theta_k \, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} \right)$$

$$\times \prod_{k=1}^{m} \theta_k^{\sum_{j=1}^{n} y_{jk}} \, exp\left(\sum_{j=1}^{n} \sum_{k=1}^{m} y_{jk}(\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}) \right)$$

$$\times exp\left\{ -\frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} - \frac{\tau \sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}$$

$$\times \tau^{mn+a-1} e^{-\frac{\tau}{b}} p^{-\frac{mn}{2}+c-1} (1-p)^{-\frac{mn}{2}+d-1} \rho^{e-1} (1-\rho)^{f-1} (1-\rho^2)^{-\frac{m(n-1)}{2}}$$

$$\times exp\left\{ -\frac{(\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{\left(\beta - \mu_\beta\right)^2}{2\sigma_\beta^2} \right\}$$

Next we will derive the conditional distribution for each parameter:

$$P(\tau | \ldots) \propto \tau^{mn+a-1} \exp\left\{ -\frac{\tau}{b} - \frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} \right.$$

$$\left. - \frac{\tau \sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}$$

$$\sim \Gamma\left(mn + a, \left\{ \frac{1}{b} + \frac{\sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} + \frac{\sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}^{-1} \right)$$

$$p(p | \ldots) \propto p^{-\frac{mn}{2}+c-1} (1-p)^{-\frac{mn}{2}+d-1}$$

$$\times exp\left\{ -\frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} - \frac{\tau \sum_{k=1}^{m} \sum_{j=1}^{n} V_{jk}^2}{2(1-p)} \right\}$$

$$p(\rho | \ldots) \propto \rho^{e-1} (1-\rho)^{f-1} (1-\rho^2)^{-\frac{m(n-1)}{2}}$$

$$\times exp\left\{ -\frac{\tau \sum_{k=1}^{m} \left[\sum_{j=1}^{n-1} \left(U_{j,k} - \rho U_{j+1,k} \right)^2 + (1-\rho^2) U_{n,k}^2 \right]}{2p(1-\rho^2)} \right\}$$

$$p(\alpha | \ldots) \propto$$

$$exp\left(-n\sum_{j=1}^{n}\sum_{k=1}^{m}\theta_k\, exp\{\alpha r_j+\beta r_j^2+U_{jk}+V_{jk}\}+\alpha\sum_{j=1}^{n}\sum_{k=1}^{m}y_{jk}r_j-\frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2}\right)$$

$$p(\boldsymbol{\beta}|\ldots)\propto$$

$$exp\left(-n\sum_{j=1}^{n}\sum_{k=1}^{m}\theta_k\, exp\{\alpha r_j+\beta r_j^2+U_{jk}+V_{jk}\}+\beta\sum_{j=1}^{n}\sum_{k=1}^{m}y_{jk}r_j^2-\frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right)$$

For the k th sample,

$$P(\theta_k|\ldots)\propto exp\left(-n\theta_k\sum_{j=1}^{n}exp\{\alpha r_j+\beta r_j^2+U_{jk}+V_{jk}\}\right)\theta_k^{\sum_{j=1}^{n}y_{jk}}$$

$$\sim\Gamma\left(\sum_{j=1}^{n}y_{jk}+1,\left[n\sum_{j=1}^{n}exp\{\alpha r_j+\beta r_j^2+U_{jk}+V_{jk}\}\right]^{-1}\right)$$

For the k th sample, when $j=1$:

$$P(U_{1k}|\ldots)\propto exp\left(-n\theta_k\, exp\{\alpha r_1+\beta r_1^2+U_{1k}+V_{1k}\}+U_{1k}y_{1k}-\frac{\tau(U_{1k}-\rho U_{2k})^2}{2p(1-\rho^2)}\right)$$

For $2\le j\le n-1$:

$$P\left(U_{jk}\middle|\ldots\right)\propto exp\Bigg(-n\theta_k\, exp\{\alpha r_j+\beta r_j^2+U_{jk}+V_{jk}\}+U_{jk}y_{jk}$$

$$-\frac{\tau\left[\left(U_{j-1,k}-\rho U_{j,k}\right)^2+\left(U_{j,k}-\rho U_{j+1,k}\right)^2\right]}{2p(1-\rho^2)}\Bigg)$$

when $j=n$:

$$P(U_{nk}|\ldots)\propto exp\Bigg(-n\theta_k\, exp\{\alpha r_n+\beta r_n^2+U_{nk}+V_{nk}\}+U_{nk}y_{nk}$$

$$-\frac{\tau\left(U_{n-1,k}-\rho U_{2,k}\right)^2}{2p(1-\rho^2)}\Bigg)$$

For j th base in the k th sample,

$$P(V_{jk}|\dots) \propto exp\left(-n\theta_k\, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} + V_{jk}y_{jk} - \frac{\tau V_{jk}^2}{2(1-p)}\right)$$

In each iteration of Gibbs Sampler, we will go through τ, p, ρ, α, β, $\theta_k (1 \le k \le m)$, $U_{jk} (1 \le k \le m, 1 \le j \le n)$ and $V_{jk} (1 \le k \le m, 1 \le j \le n)$, and sample each parameter from its conditional distribution. For $\theta_k (1 \le k \le m)$ and τ, we can directly sample from Gamma distribution. For α, β, $U_{jk} (1 \le k \le m, 1 \le j \le n)$ and $V_{jk} (1 \le k \le m, 1 \le j \le n)$, we use adaptive rejection sample (ARS) since their conditional distributions are log-concave. For p and ρ, we use adaptive rejection Metropolis sample (ARMS) since their conditional distributions are close to log-concave.

4.6.1.2 ICAR spatial Poisson regression model

An alternative is to use ICAR prior for the random effects U_{jk}, and then the model is referred as the ICAR model.

For the k th sample, the ICAR prior for U_{jk} is:

$$P\left(\vec{U}_{\cdot k}\middle|\sigma_u^2, \rho\right) \propto (\sigma_u^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2\sigma_u^2}\right\}$$

$$= p^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau \sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p}\right\}$$

$$P\left(\vec{V}_{\cdot k}\middle|\sigma_v^2\right) \propto (\sigma_v^2)^{-\frac{n}{2}} exp\left\{-\frac{\sum_{j=1}^{n} V_{jk}^2}{2\sigma_v^2}\right\} = (1-p)^{-\frac{n}{2}} \tau^{\frac{n}{2}} exp\left\{-\frac{\tau \sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

Then combine information in all m samples,

$$P\left(\vec{U}\middle|\sigma_u^2, \rho\right) \propto \prod_{k=1}^{m} P\left(\vec{U}_{\cdot k}\middle|\sigma_u^2, \rho\right) = p^{-\frac{mn}{2}} \tau^{\frac{mn}{2}} \exp\left\{-\frac{\tau \sum_{k=1}^{m} \sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p}\right\}$$

$$P(\vec{V}|\sigma_v^2) = \prod_{k=1}^{m} P(\vec{V}_{\cdot k}|\sigma_v^2) \propto (1-p)^{-\frac{mn}{2}} \tau^{\frac{mn}{2}} exp\left\{-\frac{\tau \sum_{k=1}^{m}\sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

The posterior probability is:

$$P(\vec{\theta}, \vec{U}, \vec{V}, \tau, p, \alpha, \beta|\vec{Y}) \propto P(\vec{Y}|\vec{\theta}, \vec{U}, \vec{V}, \alpha, \beta) P(\vec{U}|\tau, p) P(\vec{V}|\tau, p) P(\tau) P(p) P(\vec{\theta}) p(\alpha) p(\beta)$$

$$= exp\left(-n \sum_{j=1}^{n}\sum_{k=1}^{m} \theta_k \, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\}\right)$$

$$\times \prod_{k=1}^{m} \theta_k^{\sum_{j=1}^{n} y_{jk}} \, exp\left(\sum_{j=1}^{n}\sum_{k=1}^{m} y_{jk}(\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk})\right)$$

$$\times exp\left\{-\frac{\tau \sum_{k=1}^{m}\sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p} - \frac{\tau \sum_{k=1}^{m}\sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

$$\times \tau^{mn+a-1} e^{-\frac{\tau}{b}} p^{-\frac{mn}{2}+c-1} (1-p)^{-\frac{mn}{2}+d-1} exp\left\{-\frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right\}$$

Next we will derive the conditional distribution for each parameter:

$$P(\tau|\ldots) \propto \tau^{mn+a-1} \exp\left\{-\frac{\tau}{b} - \frac{\tau \sum_{k=1}^{m}\sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p} - \frac{\tau \sum_{k=1}^{m}\sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

$$\sim \Gamma\left(mn + a, \left\{\frac{1}{b} + \frac{\sum_{k=1}^{m}\sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p} + \frac{\sum_{k=1}^{m}\sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}^{-1}\right)$$

$$p(p|\ldots) \propto p^{-\frac{mn}{2}+c-1} (1-p)^{-\frac{mn}{2}+d-1}$$

$$\times exp\left\{-\frac{\tau \sum_{k=1}^{m}\sum_{j_1 \sim j_2}\left(U_{j_1,k} - U_{j_2,k}\right)^2}{2p} - \frac{\tau \sum_{k=1}^{m}\sum_{j=1}^{n} V_{jk}^2}{2(1-p)}\right\}$$

$$p(\alpha|\ldots) \propto exp\left(-n\sum_{j=1}^{n}\sum_{k=1}^{m}\theta_k\, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} + \alpha\sum_{j=1}^{n}\sum_{k=1}^{m} y_{jk} r_j - \frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2}\right)$$

$$p(\beta|\ldots) \propto exp\left(-n\sum_{j=1}^{n}\sum_{k=1}^{m}\theta_k\, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} + \beta\sum_{j=1}^{n}\sum_{k=1}^{m} y_{jk} r_j^2 - \frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right)$$

For the k th sample,

$$P(\theta_k|\ldots) \propto exp\left(-n\theta_k\sum_{j=1}^{n} exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\}\right)\theta_k^{\sum_{j=1}^{n} y_{jk}}$$

$$\sim\Gamma\left(\sum_{j=1}^{n} y_{jk} + 1, \left[n\sum_{j=1}^{n} exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\}\right]^{-1}\right)$$

For the k th sample, when $j = 1$:

$$P(U_{1k}|\ldots) \propto exp\left(-n\theta_k\, exp\{\alpha r_1 + \beta r_1^2 + U_{1k} + V_{1k}\} + U_{1k} y_{1k} - \frac{\tau(U_{1k} - U_{2k})^2}{2p}\right)$$

For $2 \le j \le n-1$:

$$P(U_{jk}|\ldots) \propto exp\left(-n\theta_k\, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} + U_{jk} y_{jk} - \frac{\tau\left[(U_{j-1,k} - U_{j,k})^2 + (U_{j,k} - U_{j+1,k})^2\right]}{2p}\right)$$

when $j = n$:

$$P(U_{nk}|\dots) \propto exp\left(-n\theta_k\, exp\{\alpha r_n + \beta r_n^2 + U_{nk} + V_{nk}\} + U_{nk}y_{nk} - \frac{\tau\left(U_{n-1,k} - U_{2,k}\right)^2}{2p}\right)$$

For j th base in the k th sample,

$$P(V_{jk}|\dots) \propto exp\left(-n\theta_k\, exp\{\alpha r_j + \beta r_j^2 + U_{jk} + V_{jk}\} + V_{jk}y_{jk} - \frac{\tau V_{jk}^2}{2(1-p)}\right)$$

In each iteration of the Gibbs sampler, we go through τ, p, α, β, $\theta_k(1 \le k \le m)$, $U_{jk}(1 \le k \le m, 1 \le j \le n)$ and $V_{jk}(1 \le k \le m, 1 \le j \le n)$, and sample each parameter from its conditional distribution. For $\theta_k(1 \le k \le m)$ and τ, we can directly sample from a Gamma distribution. For α, β, $U_{jk}(1 \le k \le m, 1 \le j \le n)$ and $V_{jk}(1 \le k \le m, 1 \le j \le n)$, we use adaptive rejection sample (ARS) since their conditional distributions are log-concave. For p, we use adaptive rejection Metropolis sample (ARMS) since its conditional distribution is close to log-concave.

4.6.2 Performance of the normalized sample mean method

Let Y_{ijk} represent the sequencing depth at the j th base of a specified exon where $j = 1, \dots, n$, and $n = 100$ is the length of the exon; $k = 1, \dots, m$, and $m = 7$ is the number of duplicated samples. Here i is an index for the experimental conditions: $i = 1$ indicates the case sample, and $i = 0$ indicates the control sample. We simulated seven cases and seven controls from the joint spatial Poisson regression model separately:

$$Y_{ijk}|\theta_{ik}, U_{ijk}, V_{ijk} \sim Poisson\left(n\theta_{ik}\exp\{U_{ijk} + V_{ijk}\}\right)$$

$$U_{i1k}, U_{i2k}, \dots, U_{ink} \sim N(0, \Sigma_{ui}), Var\left(U_{ijk}\right) = \sigma_{ui}^2, Cor\left(U_{ij_1k}, U_{ij_2k}\right) = \rho^{|j_1 - j_2|}$$

$$V_{ijk}\ iid \sim N(0, \sigma_{vi}^2)$$

In the normalized sample mean method, the estimates of exon level gene expression are:

$$\hat{\theta}_{ik} = \frac{\sum_{j=1}^{n} Y_{ijk}}{n^2}$$

Which is an unbiased estimate of

$$\frac{EY_{ijk}}{n} = \theta_{ik} exp\left\{\frac{\sigma_{ui}^2 + \sigma_{vi}^2}{2}\right\} \geq \theta_{ik}$$

The moment estimate of fold change is:

$$\frac{E\hat{\theta}_{1k}}{E\hat{\theta}_{0k}} = \frac{\theta_{1k}}{\theta_{0k}} exp\left\{\frac{\sigma_{u1}^2 + \sigma_{v1}^2 - \sigma_{u0}^2 - \sigma_{v0}^2}{2}\right\}$$

When the data Y_{ijk} are simulated from the joint spatial Poisson model, the normalized sample mean method will always overestimate the true expression level. Larger total variance of two random effects indicates larger bias. However, the bias of fold change estimator is not simple. When the total variances of two random effects are the same in two experimental conditions ($\sigma_{u1}^2 + \sigma_{v1}^2 = \sigma_{u0}^2 + \sigma_{v0}^2$), the estimate of fold change is unbiased. When the total variance of two random effects in case is larger than that in control ($\sigma_{u1}^2 + \sigma_{v1}^2 > \sigma_{u0}^2 + \sigma_{v0}^2$), the estimate of fold change will overestimate the true value, while when the total variance of two random effects in case is smaller than that in control ($\sigma_{u1}^2 + \sigma_{v1}^2 < \sigma_{u0}^2 + \sigma_{v0}^2$), the estimate of fold change will underestimate the true value. To account for these complicated cases, we considered five different cases for variance of the two random effects (Table 4.1).

4.6.3 Preliminary analysis of a real RNA-Seq dataset

In this study we analyzed the sequencing quality control project (SEQC) dataset (unpublished data from Dr Jun Li, Department of Human Genetics, University of Michigan), which consists of two samples from microarray quality control project (MAQC) (125): brain and universal human reference (UHR). They were prepared using

the standard Illumina mRNA-Seq protocol and reagents and sequenced across seven lanes (across two single flowcells) to a depth of 35 bases. The data were generated using Illumina pipeline 1.1 and eland_rna which is optimized for alignment of transcriptomic reads. Only reads that uniquely mapped to the human reference genome, with up to two mismatches, were included in this study. Exon level annotations for 385,122 ENSEMBL exons were downloaded from ENSEMBL database version 55: http://jul2009.archive.ensembl.org/biomart/martview/ with genome assembly GRCh37. Using UCSC genome browser tool "liftover" to convert GRCh37 to NCBI36, total number of exons was reduced to 384,763, and then the base-level sequencing depth files were generated for these 384,763 exons in seven brain samples and seven UHR samples. We applied three exon filters in this study: non-zero filter, over-dispersion filter and non-gap filter. The number of exons left in the brain samples and the UHR samples are listed in Table A4.1.

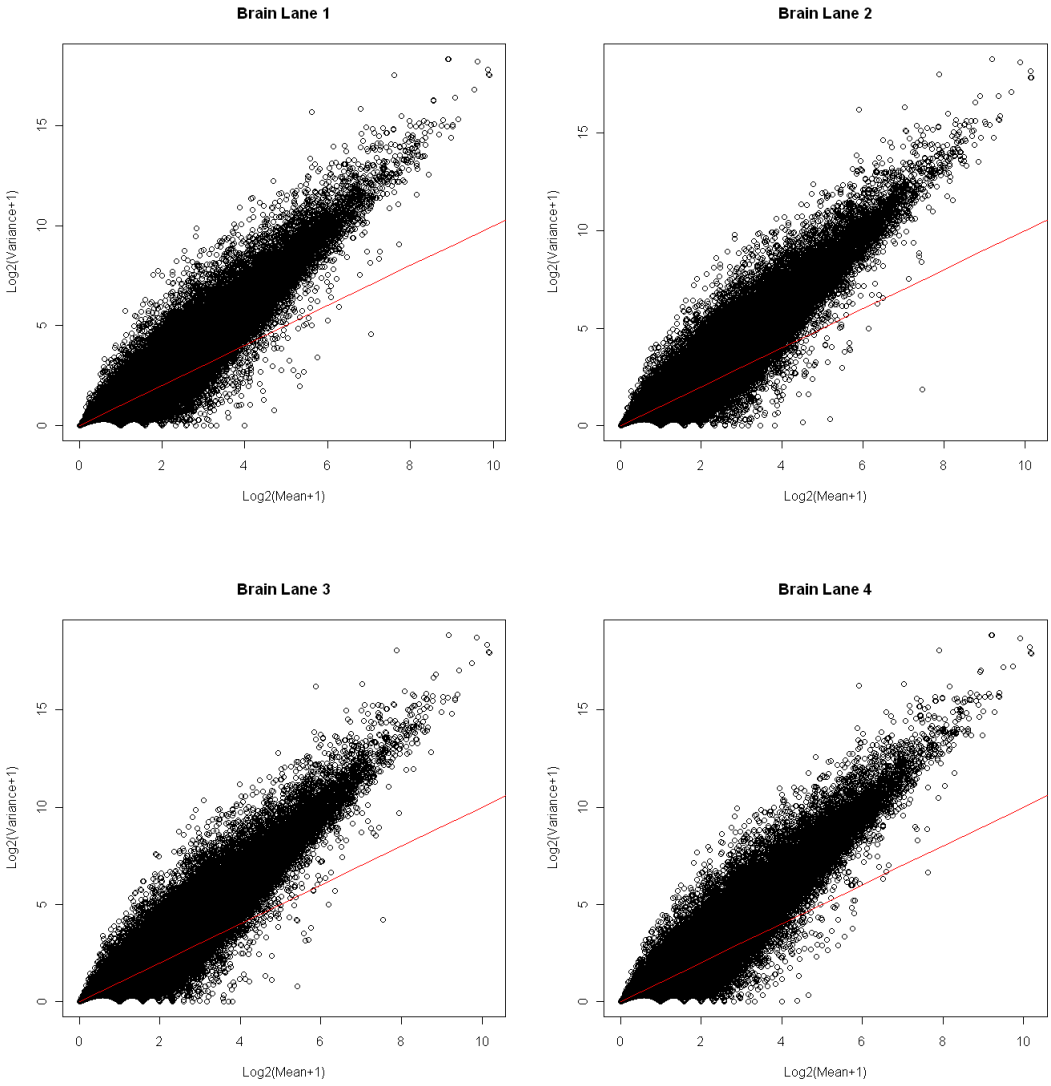
Table A4.1: The number of exons left in the brain samples and the UHR samples after three exon filters

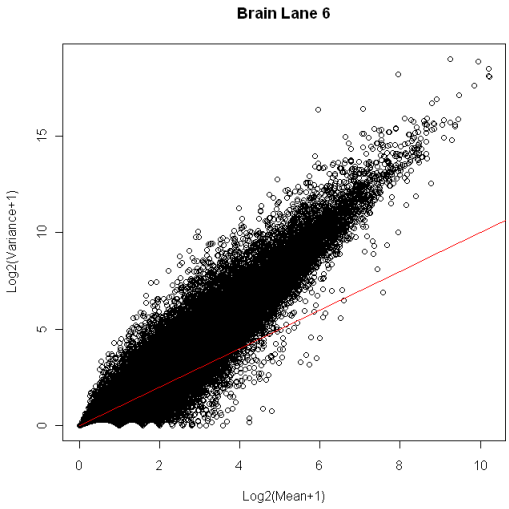
Filters	Brain	UHR
Total number of exons	384,763	384,763
After non-zero filter	170,877	185,253
After over-dispersion filter	43,141	56,947
After non-gap filter	3,569	5,241

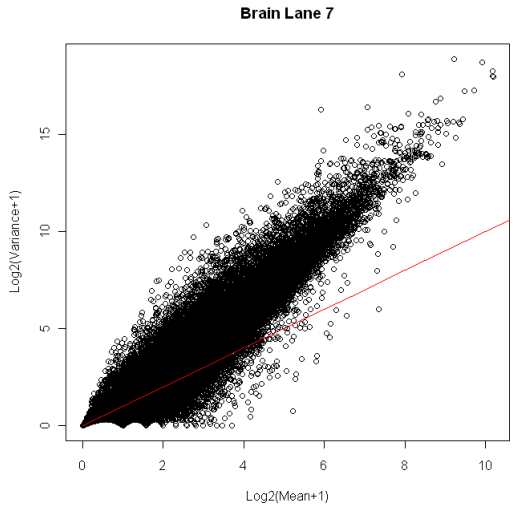
We plotted the Log2 RPKM versus the Log2 variance for 170,877 non-zero exons in seven brain samples and 185,253 non-zero exons in seven UHR samples in Figure A4.1. From Figure A4.1, we observed substantial over-dispersion: the variance of the

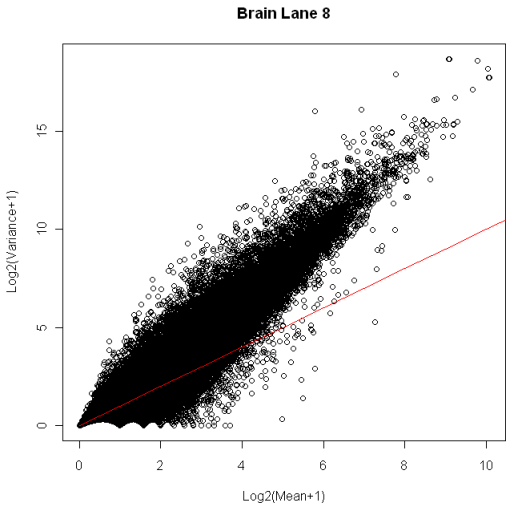
number of reads in each exon is larger than the mean of the number of reads in each exon in the majority of non-zero exons, especially for the highly expressed exons.

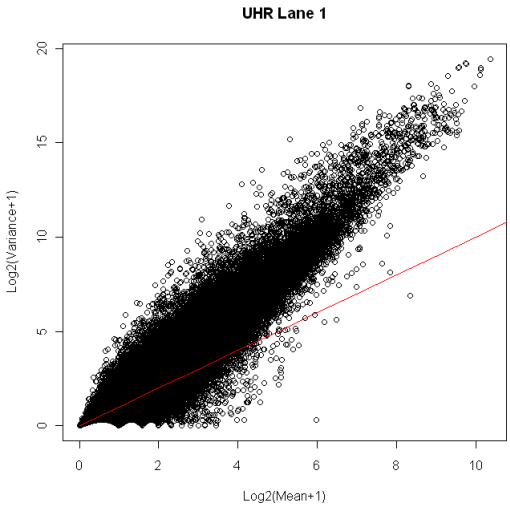
Figure A4.1: Log2 RPKM versus Log2 variance in all 14 samples. The red line is the diagonal line where the mean and the variance are the same.

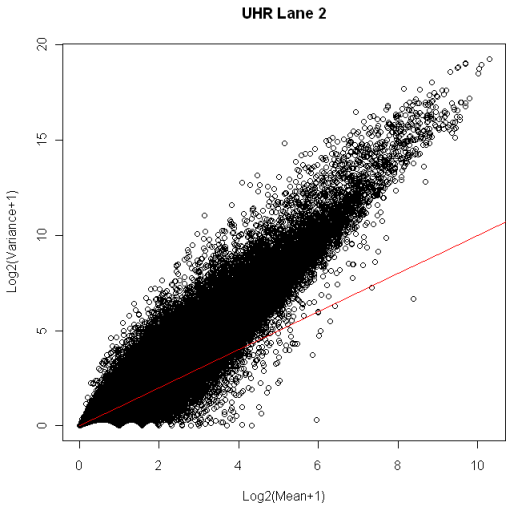


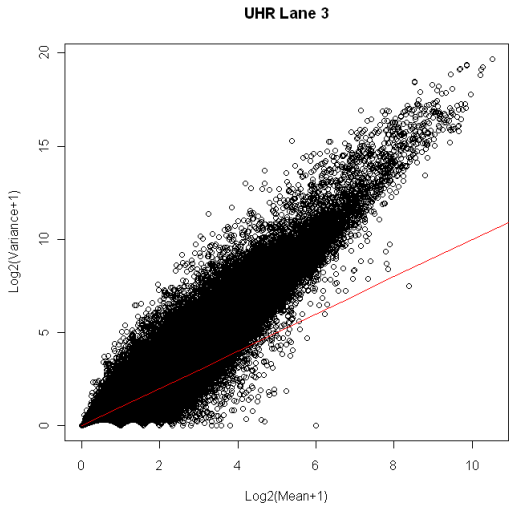

Brain Lane 6
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
0
2
4
6
8
10

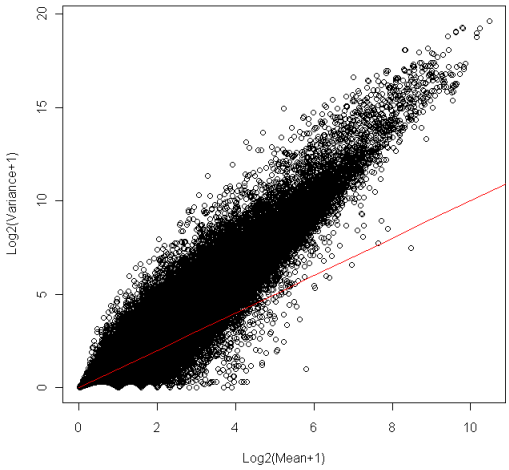

Brain Lane 7
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
0
2
4
6
8
10

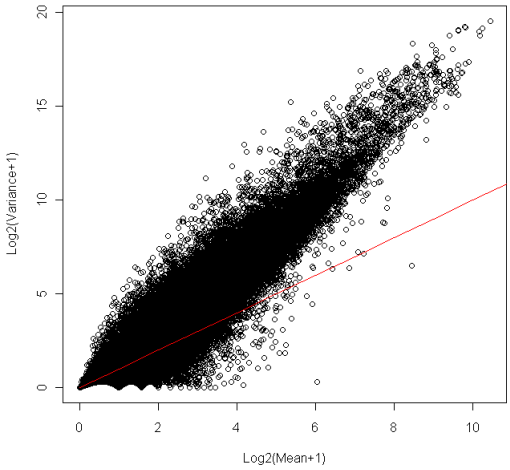

Brain Lane 8
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
0
2
4
6
8
10

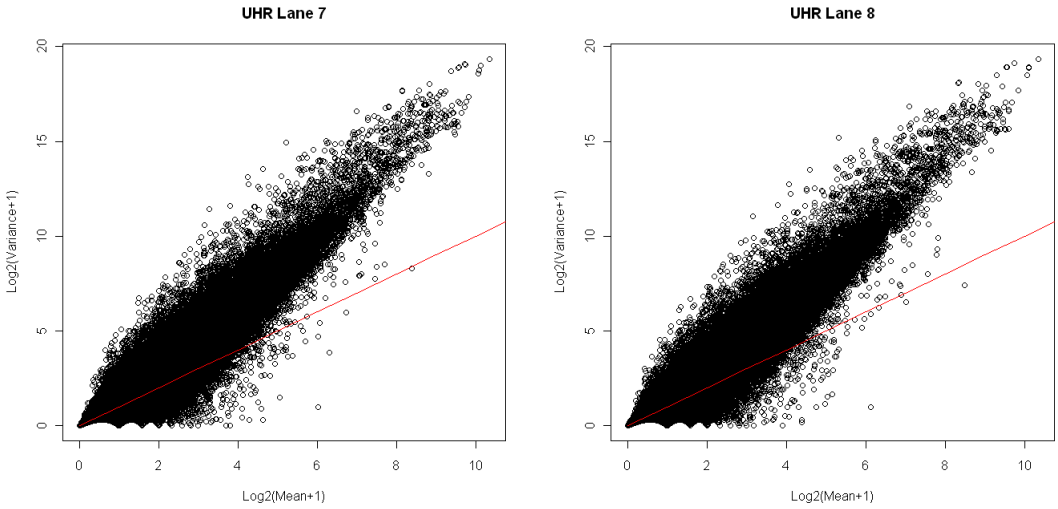

UHR Lane 1
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
20
0
2
4
6
8
10


UHR Lane 2
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
20
0
2
4
6
8
10


UHR Lane 3
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
20
0
2
4
6
8
10

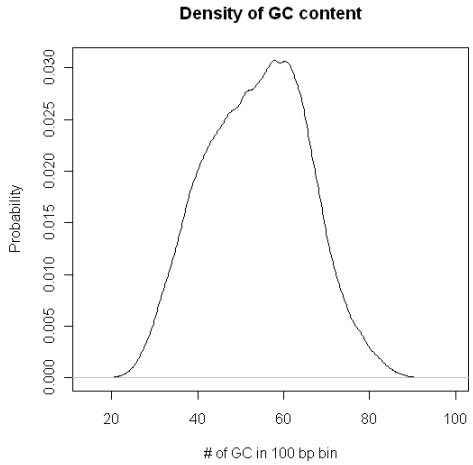

UHR Lane 4
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
20
0
2
4
6
8
10


UHR Lane 6
Log2(Variance+1)
Log2(Mean+1)
0
5
10
15
20
0
2
4
6
8
10



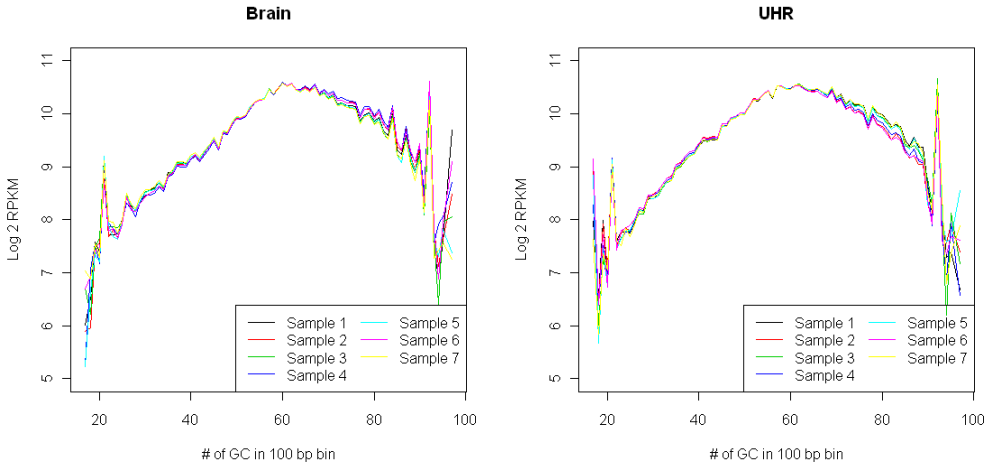
Next we explored the correlation between GC content and sequencing depth. We divided each of these 384,763 exons into 100 bp bins, and counted the number of "G" and "C" in each bin. The GC content is defined as the sum of the "G" count and the "C" count. The average length of these 384,763 exons is 293 bp. We removed the bins with length less than 100 bp, and then got 912,929 100 bp bins. Then we calculated the number of mapped reads in each 100 bp bin, and removed the bins with zero read coverage. The total number of 100 bp bins was reduced to 297,404. The distribution of GC content in these 297,404 100 bp bins looked similar to normal distribution (Figure A4.2): with range from 17 to 97, with mean 53.73 and standard deviation 11.97.

Figure A4.2: Distribution of GC content in 297,404 100 bp bins



Next we calculated the Log2 RPKM (reads per thousand base pair per million uniquely mapped reads) for each of these 297,404 100 bp bins in each sample, and then took the average of Log2 RPKM across the 100 bp bins with the same GC content. Figure A4.3 shows the quadratic patterns between GC content and Log2 RPKM in seven brain samples and seven UHR samples.

Figure A4.3: Quadratic patterns between GC content and Log2 RPKM in seven brain samples and seven UHR samples.



From Figure A4.3, we observed a quadratic pattern between GC content and Log2 RPKM: the bins with small or large GC content have lower sequencing depth than the bins with median GC content, except for the noise at the two edges (number of GC < 20 and numebr of GC > 90) which may due to lack of sample size. In addition, the within sample variation is very small, but there may exist some between sample variation.

Next we took average of Log2 RPKM with same GC content across seven duplicated samples in brain or UHR, and then fitted a simple weighted linear regression with quadratic term. The weight is the frequency of GC content (Figure A4.2). We assumed that the bin with GC content 50 has Log2 RPKM zero. The normalized GC% was defined as:

$$Normalized\ GC\% = \frac{\#\ of\ GC\ in\ 100\ bp\ bin - 50}{100}$$

Figure A4.4 shows the raw data and the fitted lines from the simple linear regression.

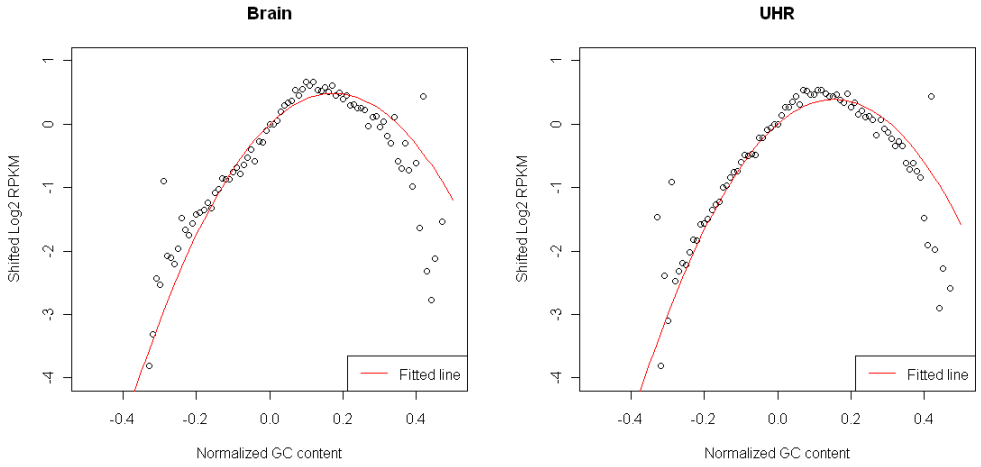
For brain samples:

$$Shifted\ Log2\ RPKM = 5.5556 * Normalized\ GC\% - 15.8962 * Normalized\ GC\%^2$$

For UHR samples:

$$Shifted\ Log2\ RPKM = 5.0531 * Normalized\ GC\% - 16.4003 * Normalized\ GC\%^2$$

Figure A4.4: Simple weighted linear regression with quadratic term



To incorporate GC content as covariates, we add both linear and quadratic term of GC conent:

$$Y_{jk}|\theta_k, U_{jk}, V_{jk}, \alpha, \beta \sim Poisson\left(n\theta_k \exp\left\{\alpha * (r_j - \bar{r}) + \beta * (r_j - \bar{r})^2 + U_{jk} + V_{jk}\right\}\right)$$

Where $\alpha * (r_j - \bar{r}) + \beta * (r_j - \bar{r})^2$ represent the fixed effect of GC content, U_{jk}'s are the random effect for the spatial correlation, and V_{jk}'s are the random effect for unstructured variability. Assume the reads is 35 bp. We extend the j th base to both sides for 34 bp to make a 69 bp bin, and then define r_j as the percentage of "G" and "C" in this 69 bp bin centered at the j th base. $\bar{r}$ is defined as the mean percentage of "G" and "C" in the specified exon. θ_k is the exon level expression index for the k th sample with GC content $\bar{r}$, which is the parameter of interest. Note that we observed very little within sample variation, therefore the parameters α and β are the same across all m duplicated samples.

We draw trace plots (Figure A4.5) for the exon ENSE00001701801 (chr 17, 1250091~1250306) from seven brain samples. It contains 216 bases, and shows substantial over-dispersion and spatial correlation (Table A4.2).

Figure A4.5: Trace plots for the exon ENSE00001701801

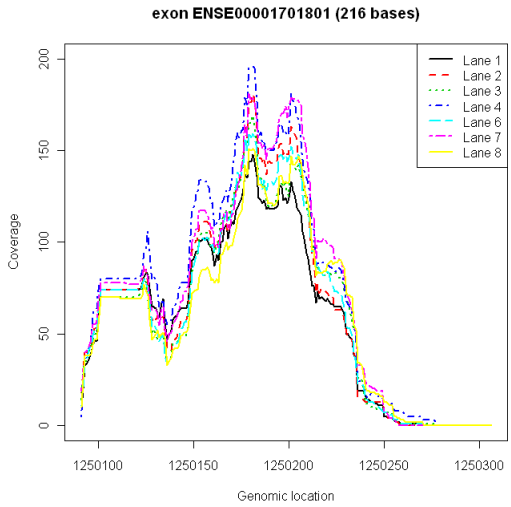


Table A4.2: Mean, variance and ACF for exon ENSE00001701801 in seven brain samples

Lane	Mean	Variance	ACF(lag=1)	ACF(lag=20)
1	58	2068	0.9893	0.7095
2	64	2852	0.9903	0.6811
3	61	2489	0.9896	0.6736
4	74	3417	0.9878	0.6967
6	62	2581	0.9904	0.6773
7	71	3258	0.9902	0.6867
8	60	2190	0.9884	0.6690

CHAPTER 5

Conclusion

This thesis presented model-based methods for the analysis of high throughput genomic data. In the presence of technologically specific features, heterogeneous data structures and massive sample sizes, Bayesian approaches can be a powerful modeling framework. The research in this dissertation demonstrates that Bayesian modeling approaches have achieved great success and have the potential to accelerate biomedical research.

Chapter 2 describes a model-based Bayesian variable selection approach to query large scale microarray compendium datasets. By modeling the observed microarray compendium data as a mixture of normal distributions, this approach identifies transcript factor target genes under a subset of experimental conditions. Further, it is capable of detecting complicated co-expression patterns, such as inversed patterns, and it is robust in the presence of sporadic outliers. All these unique features are critical to improve the power of differentiating biological signals from background noise. Due to the high dimensionality of unknown parameters, the commonly used maximum likelihood

estimate is not feasible. In this context, a Bayesian modeling approach coupled with Markov Chain Monte Carlo estimation techniques appears to be an attractive alternative.

Chapter 3 presents hybrid motif sampler (HMS), a *de novo* motif finding algorithm designed for analyzing ChIP-Seq data. We model the intra-motif dependency using high dimensional multinomial distributions. The increased number of model parameters is supported by the dramatically increased sample size in ChIP-Seq experiments and describes the underlying motif pattern more accurately. Borrowing information from the base level sequencing depth data, we adopt an informative prior distribution to facilitate motif detection. We also combine stochastic sampling and a deterministic search to speed up the computationally intensive iterative procedure. Using a Bayesian modeling approach to analyze ChIP-Seq data thoroughly and efficiently will significantly improve the accuracy of existing transcript factor binding sites.

Chapter 4 introduces a spatial Poisson regression model for exon level RNA-Seq data. We utilize two random effects to explain the spatial correlation and the non-spatial variation. These two random effects are not directly observable. We need to use high dimensional integration to integrate them out in frequentist methods. Bayesian modeling with Markov chain Monte Carlo techniques provides an appealing solution by iteratively sampling each random effect from conditional distributions. We also incorporate GC content effects into the mean structure for better fitting, and it can be easily extended to adjust for other biological effects and technical effects. The analysis of the SEQC dataset shows that the spatial Poisson regression model has the ability to improve quantification of the true underlying expression levels.

In all chapters, the Bayesian modeling framework serves as the consistent procedure for combining information from high throughput genomic data and prior information. The fast and efficient Markov Chain Monte Carlo techniques allow statistical inference despite of the high dimensionality of unknown parameters. With the development of new biological technologies, scientists are facing an analytic bottleneck of accumulating high throughput genomic data. Therefore, Bayesian modeling approaches will be of greatest benefit to the broader biomedical community for their simplicity, efficiency and practical advantages.

REFERENCES

1. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y*, **270**, 467-470.
2. Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349-360.
3. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews*, **10**, 57-63.
4. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y*, **316**, 1497-1502.
5. Kauffman, S., Peterson, C., Samuelsson, B. and Troein, C. (2003) Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14796-14799.
6. Lee, I., Date, S.V., Adai, A.T. and Marcotte, E.M. (2004) A probabilistic functional network of yeast genes. *Science (New York, N.Y*, **306**, 1555-1558.
7. Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M. (2003) Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics (Oxford, England)*, **19**, 643-650.
8. Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian data analysis*. Reprinted 1997. ed. Chapman & Hall, London.
9. Hu, M. and Qin, Z.S. (2009) Query large scale microarray compendium datasets using a model-based bayesian approach with variable selection. *PLoS ONE*, **4**, e4495.

10. Hu, M., Yu, J., Taylor, J.M., Chinnaiyan, A.M. and Qin, Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research*, **38**, 2154-2167.
11. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**, 1675-1680.
12. Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet*, **21**, 33-37.
13. Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, **5**, e8.
14. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.
15. Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. and Davidson, G.S. (2001) A gene expression map for Caenorhabditis elegans. *Science (New York, N.Y.*, **293**, 2087-2092.
16. Hibbs, M.A., Hess, D.C., Myers, C.L., Huttenhower, C., Li, K. and Troyanskaya, O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics (Oxford, England)*, **23**, 2692-2699.
17. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863-14868.
18. Bassett, D.E., Jr., Eisen, M.B. and Boguski, M.S. (1999) Gene expression informatics--it's all in your mine. *Nat Genet*, **21**, 51-55.
19. Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D. and Klingler, T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome research*, **9**, 1198-1203.
20. Quackenbush, J. (2003) Genomics. Microarrays--guilt by association. *Science (New York, N.Y.*, **302**, 240-241.

21. Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, **6**, 227.
22. Mootha, V.K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., Delmonte, T., Villeneuve, A., Sladek, R., Xu, F. *et al.* (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 605-610.
23. Butte, A.J. and Kohane, I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418-429.
24. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.
25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
26. Neuwald, A.F., Kannan, N., Poleksic, A., Hata, N. and Liu, J.S. (2003) Ran's C-terminal, Basic Patch, and Nucleotide Exchange Mechanisms in Light of a Canonical Structure for Rab, Rho, Ras, and Ran GTPases. *Genome Res*, **13**, 673-692.
27. Tanay, A., Sharan, R. and Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics (Oxford, England)*, **18 Suppl 1**, S136-144.
28. Cheng, Y. and Church, G.M. (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, **8**, 93-103.
29. Getz, G., Levine, E. and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12079-12084.
30. Sheng, Q., Moreau, Y. and De Moor, B. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics (Oxford, England)*, **19 Suppl 2**, II196-II205.
31. Madeira, S.C. and Oliveira, A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **1**, 24-45.

32. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. and Gerstein, M. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, **314**, 1053-1066.
33. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
34. Dhollander, T., Sheng, Q., Lemmens, K., Moor, B.D., Marchal, K. and Moreau, Y. (2007) Query-driven module discovery in microarray data. *Bioinformatics (Oxford, England)*.
35. Owen, A.B., Stuart, J., Mach, K., Villeneuve, A.M. and Kim, S. (2003) A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome research*, **13**, 1828-1837.
36. Chen, R. and Liu, J.S. (1996) Predictive Updating Methods With Application to Bayesian Classification. *Journal of the Royal Statistical Society Series B-Methodological*, **58**, 397-415.
37. Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.
38. Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. springer Verlag, Berlin.
39. Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J. and Gardner, T.S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic acids research*, **36**, D866-870.
40. Brinkman, A.B., Ettema, T.J., de Vos, W.M. and van der Oost, J. (2003) The Lrp family of transcriptional regulators. *Molecular microbiology*, **48**, 287-294.
41. Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J. *et al.* (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic acids research*, **34**, D394-397.
42. Urbanowski, M.L., Stauffer, L.T. and Stauffer, G.V. (2000) The gcvB gene encodes a small untranslated RNA involved in expression of the dipeptide and

oligopeptide transport systems in Escherichia coli. *Molecular microbiology*, **37**, 856-868.

43. Tani, T.H., Khodursky, A., Blumenthal, R.M., Brown, P.O. and Matthews, R.G. (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13471-13476.
44. Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, **16**, 939-945.
45. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907-2912.
46. McLachlan, G.J., Bean, R.W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics (Oxford, England)*, **18**, 413-422.
47. Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics (Oxford, England)*, **17**, 763-774.
48. Ghosh, D. and Chinnaiyan, A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics (Oxford, England)*, **18**, 275-286.
49. Medvedovic, M. and Sivaganesan, S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics (Oxford, England)*, **18**, 1194-1206.
50. Qin, Z.S. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics (Oxford, England)*, **22**, 1988-1997.
51. Kim, S., Tadesse, M.G., Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**, 877-893.
52. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, **18**, 6097-6100.
53. Bembom, O. (2007) Sequence logos for DNA sequence alignments.

54. Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, **20**, 835-839.
55. Liu, X., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-138.
56. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting Subtle Sequence Signals : a Gibbs Sampling Strategy For Multiple Alignment. *Science (New York, N.Y.*, **262**, 208-214.
57. Bailey, T.L. and Elkan, C.P. (1994) Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. *ISMB*, 28-36.
58. Bussemaker, H.J., Li, H. and Siggia, E.D. (2000) Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Nat'l Acad. Sci. USA*, **97**, 10096--10100.
59. Stormo, G.D. and Hartzell III, G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the Nathional Academy of Science, USA*, **86**, 1183-1187.
60. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, **23**, 137-144.
61. McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. and Lawrence, C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic acids research*, **29**, 774-782.
62. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27**, 167-171.
63. Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3339-3344.
64. Shim, H. and Keles, S. (2008) Integrating quantitative information from ChIP-chip experiments into motif finding. *Biostatistics*, **9**, 51-65.
65. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. *et al.* (2000) Genome-wide

location and function of DNA binding proteins. *Science (New York, N.Y.*, **290**, 2306-2309.

66. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533-538.
67. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.
68. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, **4**, 651-657.
69. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553-560.
70. Solomon, M.J., Larsen, P.L. and Varshavsky, A. (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937-947.
71. Orlando, V. and Paro, R. (1993) Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell*, **75**, 1187-1198.
72. Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M. and Jones, S.J. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics (Oxford, England)*, **24**, 1729-1730.
73. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.
74. Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*.

75. Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research*, **36o**, 5221-5231.
76. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M. and Wong, W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, **26**, 1293-1300.
77. Kharchenko, P.V., Tolstorukov, M.Y. and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, **26**, 1351-1359.
78. Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, **27**, 66-75.
79. Choi, H., Nesvizhskii, A.I., Ghosh, D. and Qin, Z.S. (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics (Oxford, England)*, **25**, 1715-1721.
80. Nix, D.A., Courdy, S.J. and Boucher, K.M. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, **9**, 523.
81. Lawrence, C.E. and Reilly, A.A. (1990) An expectation maximization (EM) algorithm for the idenification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41-51.
82. Liu, J.S. (1994) The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene-Regulation Problem. *Journal of the American Statistical Association*, **89**, 958-966.
83. Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1995) Bayesian Models For Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Assoc*, **90**, 1156-1170.
84. Neuwald, A.F., Liu, J.S. and Lawrence, C.E. (1995) Gibbs Motif Sampling : Detection of Bacterial Outer-Membrane Protein Repeats. *Protein Science*, **4**, 1618-1632.
85. Gupta, M. and Liu, J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. *Journal of the American Statistical Association*, **98**, 55-66.

86. Staden, R. (1988) Methods to define and locate patterns of motifs in sequences. *Comput Appl Biosci*, **4**, 53-60.
87. Bulyk, M.L., Johnson, P.L. and Church, G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, **30**, 1255-1261.
88. Lee, M.L., Bulyk, M.L., Whitmore, G.A. and Church, G.M. (2002) A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics*, **58**, 981-988.
89. Benos, P.V., Bulyk, M.L. and Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic acids research*, **30**, 4442-4451.
90. Man, T.K. and Stormo, G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic acids research*, **29**, 2471-2478.
91. King, O.D. and Roth, F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic acids research*, **31**, e116.
92. Barash, Y., Elidan, G., Friedman, N. and Kaplan, T. (2003), *RECOMB 2003*, Berlin, Germany.
93. Zhou, Q. and Liu, J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics (Oxford, England)*, **20**, 909-916.
94. 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 Sciences of the United States of America*, **98**, 5116-5121.
95. Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S., Yu, J. and Chinnaiyan, A.M. (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.
96. Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics (Oxford, England)*, **14**, 48-54.
97. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.

98. Leach, A.R. (1996) *Molecular modelling : principles and applications*. NY Longman Pub. Group, White Plains.
99. Jensen, S.T. and Liu, J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics (Oxford, England)*, **20**, 1557-1564.
100. Consul, P.C. (1989) *Generalized Poisson Distributions*. Marcel Dekker, New York.
101. Johnson, N.L., Kotz, S. and Kemp, A.W. (1992) *Univariate discrete distributions*. 2nd ed. John Wiley & Sons, New York.
102. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.*, **320**, 1344-1349.
103. Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bahler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239-1243.
104. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-628.
105. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523-536.
106. Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, **5**, 613-619.
107. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509-1517.
108. Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, **18**, 610-621.
109. Bullard, J.H., Purdom, E., Hansen, K.D. and Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, **11**, 94.

110. Jiang, H. and Wong, W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032.
111. Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31-36.
112. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, **36**, e105.
113. Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society, Ser. B*, **36**, 192-236.
114. Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.
115. Besag, J., York, J. and Mollie, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1-20.
116. Cressie, N. and Chan, N. (1989) Spatial Modeling of Regional Variables. *Journal of the American Statistical Association*, **84**, 393-401.
117. Best, N., Richardson, S. and Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**, 35-59.
118. Wakefield, J. (2007) Disease mapping and spatial regression with count data. *Biostatistics*, **8**, 158-183.
119. Green, P. and Richardson, S. (2002) Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*, **97**, 1055-1070.
120. Knorr-Held, L. and Raber, G. (2000) Bayesian Detection of Clusters and Discontinuities in Disease Maps. *Biometrics*, **56**, 13-21.
121. Diggle, P., Tawn, J. and Moyeed, R. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299-350.
122. Zhu, H., Liang, F., Gu, M. and Perterson, B. (2009) Stochastic approximation algorithms for estimation of spatial mixed models. *Manuscript*.
123. Gilks, W. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337-348.
124. Gilks, W., Best, N. and Tan, K. (1995) Adaptive rejection Metropolis sampling. *Applied Statistics*, **44**, 455-472.

125. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151-1161.
126. Dean, C.B. (1992) Testing for overdispersoin in Poisson and binomial regression models. *Journal of the American Statistical Association*, **87**, 451-457.
127. Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V. and Majewski, J. (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC genomics*, **9**, 529.
128. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research*, **33**, e175.
129. Li, J., Jiang, H. and Wong, W.H. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol*, **11**, R50.
130. Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet*, **24**, 133-141.
131. Wold, B. and Myers, R.M. (2008) Sequence census methods for functional genomics. *Nat Methods*, **5**, 19-21.
132. Asmann, Y.W., Klee, E.W., Thompson, E.A., Perez, E.A., Middha, S., Oberg, A.L., Therneau, T.M., Smith, D.I., Poland, G.A., Wieben, E.D. *et al.* (2009) 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics*, **10**, 531.
133. Gelfand, A.E., Dey, D. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling based methods (with discussion). *Bayesian Statistics 4. Eds: J. Bernardo et al.*, 147-167.