UNIVERSITY OF CALIFORNIA
RIVERSIDE

Low-Level and High-Level Microarray Data Analysis

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Xin Chen

December 2010

Dissertation Committee:

Dr. Xinping Cui, Co-Chairperson
Dr. Shizhong Xu, Co-Chairperson
Dr. Daniel Jeske

The Dissertation of Xin Chen is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

Acknowledgments

First of all, I would like to express my gratitude to my advisor, Dr. Xinping Cui for her guidance, support, encouragement, and patience during my graduate studies. You are the first person leading me into amazing biostatistics world which requires a lot of effort to learn not only statistics, but also biology, genetics, and programming. Dr. Cui has given me guidance in all these subjects and encouraged me to develop independent thinking and research skills as well.

Also I really appreciate my co-advisor, Dr. Shizhong Xu. Your support and invaluable advice gave me a lot of strength and inspiration. Your passion on research touched me deeply.

Many thanks to Dr. Daniel Jeske to be my committee member. Your consulting class was one of the best classes I have ever taken.

I owe my thanks to my friends in Dr Cuis research group, Gabriel Murillo, Nigie Shi, Bushi Wang, Na You, Jason Wilson, Zhanpan Zhang, and Haibing Zhao for the discussions, suggestions and encouragement.

Finally, I thank my parents, to whom I dedicate the dissertation, and my soprano kitty, for her support and encouragement.

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Low-Level and High-Level Microarray Data Analysis

by

Xin Chen

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2010
Dr. Xinping Cui, Co-Chairperson
Dr. Shizhong Xu, Co-Chairperson

Microarray data analysis involves low-level and high-level analysis. The low-level analysis focuses on how to get accurate and precise gene expression data. The analysis built on gene expression data is the high-level analysis such as differential gene expression analysis, SFP detection, eQTL analysis and so on. This thesis focuses on applications in both low-level and high-level analysis. In the low-level analysis, the proposed L-GCRMA method combines the advantage of the GCRMA model and the Langmuir model to get a more accurate and precise gene expression data, especially at high concentration. The simulation study and spike-in data analysis demonstrates the advantage of proposed L-GCRMA model. In the high-level analysis, a well developed SEM algorithm is successfully applied to eQTL analysis and trait-gene association analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In past decade, Affymetrix GeneChip arrays have rapidly become the most popular tool for large scale gene expression analysis in many areas of biological and medical research.In this chapter, we will first briefly review DNA structure and gene expression, followed by the introduction of Microarray technology. Finally we will review statistical methodology involved in Microarray study.

1.1 DNA and Gene Expression

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. The basic building block of DNA is nucleotide which consists three components: a phosphate group, a deoxyribose sugar and a base. There are four different bases: adenine, cytosine, guanine and thymine known by letters A, C, G, T. Discovered by Wilkins et al. (1953), a DNA molecule consists of a double helix held together using hydrogen bonding. Bases A and T or G and C are referred as complimentary bases because hydrogen bonds can form between the A-T or G-T. In the process of gene expression, DNA need to be copied and

transcribed into the form of ribonucleic acid (RNA) molecules, which is single-stranded and complementary to one of the two DNA strands.

Ribonucleic acid (RNA) is very similar to DNA, but differs in a few important structural details: in the cell, RNA is usually single-stranded, while DNA is usually double-stranded; RNA has the base uracil (U) rather than thymine (T) that is present in DNA. There are several types of RNA: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). The basic function of RNA is to carry out information from DNA and synthesize protein molecules based on these information.

A gene is a stretch of DNA that codes for a type of protein. It determines when, what amount and what kind of protein will be generated in the cell. The protein in turn controls a physical trait of the cell. The process of synthesizing proteins from genes is called gene expression which occurs in two stages: transcription and translation. In transcription stage, the information from double-stranded DNA is transferred to single-stranded mRNA. The translation is the process of translating the mRNA into a protein. The study of gene expression will help us better understand how these genes affect the function of cells. Traditionally, gene expression studies were done one gene at a time using technologies such as RT-PCR and Northern blots. The more recent development of microarray technologies allows the simultaneous measurement of the expression level of thousands of genes.

1.2 Microarray Technology

A DNA microarray is a multiplex technology used in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing a specific DNA sequence, known as probes. This can

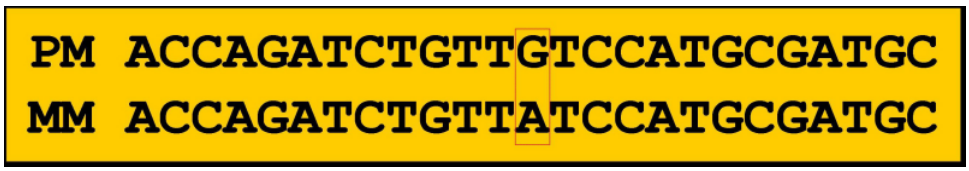

Figure 1.1: Perfect Match and Mismatch Probe

be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. There are two main microarray technologies: spotted microarray (cDNA spotted microarray and oligonucleotide spotted microarray) and in-situ oligonucleotide microarray. We only discuss oligonucleotide microarray in this dissertation.

The oligonucleotide microarray technology uses hybridization probes, which comprises hundreds of thousands of 25-mer oligonucleotide chemically synthesized on a grided array. Typically, these probes (or features) are grouped into different probe sets for different target genes. A probe set consists of 11-20 probe pair depending on different species, each of which is designed to probe a different 25 based sequence of a given gene. There are two types of probe on a chip. One is called perfect match (PM) probe which contains the exact sequence of that gene and the other is mismatch (MM) probe which is identical to the PM probe except that the middle (13th) base is converted to its complement according to Watson-Crick base pairing (see Affymetrix (2001) for details). Figure 1.1 is a example of PM and MM probes. In theory, the MM probes can be used to quantify and remove background noise. A PM and its corresponding MM probe are referred to as a probe pair.

The purpose of oligonucleotide microarray is to measure gene expression values for

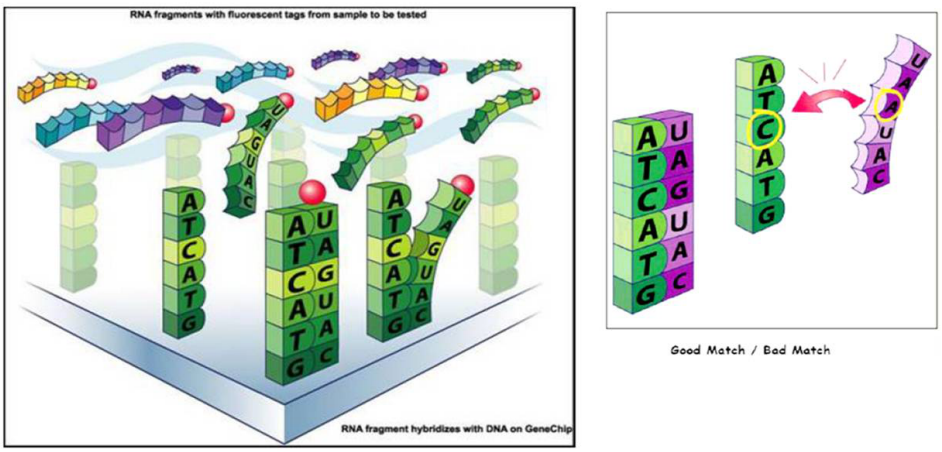

Figure 1.2: RNA Fragment Hybridizes with Probes on Oligonucleotide Array

target samples. The preparation of target samples is complex. Usually target samples contain a lot of small biotin labeled RNA fragments. The process of hybridization is very complicated. The complimentary target RNA hybridizing with a probe is considered as a good match. While it is also possible that non-complimentary target RNA hybridizes with the probe, which is a bad match. Figure 1.2 shows the procedure of hybridization. After labeled RNA samples are hybridized with arrays, images of hybridization signals will be produced and processed to obtain an intensity value for each probe. Hence, a total of 22-40 intensity values are obtained for each gene and a single composite index representing expression level of that gene must be derived.

1.3 Statistical Application in Microarray Data

1.3.1 Low-level analysis

The low-level analysis of Microarray data is also called preprocessing. The basic purpose of this analysis is to get a more biologically meaningful gene expression value

obtained from raw probe-level intensity values. Totally there are three steps in preprocessing – background correction, normalization and summarization.

The raw intensity data contains various sources of noise that are not due to biological reasons. The noise can come from the hybridization of labeled targets with probes which are not complementary to each other. This noise is also known as non-specific binding which is the main noise needed to be removed. Other noises include deposits left after the wash stage and the fluorescent intensity from the surface of chip instead of labeled strands also known as optical noise. In order to remove these noises, we need to consider a background correction method.

In multiple array analysis, the scale of different array might be different. The difference is often caused by systematic bias instead of biological variation. For example, different scanner settings might lead to different intensity readings from same samples. Normalization is the process of removing unwanted non-biological variation that might exist between arrays in a microarray experiment.

The microarray raw intensity data may be too large or meaningless for biologists who are interested in gene function instead of probe level data. Summarization is the process of combining the multiple probe intensities of each probeset to produce one expression value.

The need for preprocessing is widely acknowledged and most microarray products come with preprocessing software. In chapter 2, we propose an adjusted GCRMA preprocessing method using non-linear model in summarization.

1.3.2 Differential expression and cluster analysis

When we get preprocessed gene expression data, we are interested in analyzing fold change of gene expression under different conditions which is called differential ex-

pression analysis. Many differential expression analysis is to identify the genes whose expression levels change between two sample groups. For example, to understand the effect of a drug we may ask which genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups. Statistical methods for such data analysis include the simple t-test (Devore and Peck, 1997), the Bayesian method combined with the t-test (Baldi and Long, 2001), SAM (significance analysis of microarrays) (Tusher et al., 2001). The other type of experiments focus on examining gene expressions in multiple conditions. For example, we need to examine the effect of salt stress on gene expression in barley roots. The control may be tissues of barley roots collected from normal salt stress. The treatment may be represented by tissues sampled from barley roots treated with different degree of salt stress. Because there are multiple levels of treatment, a traditional t-test is no longer sufficient and an analysis of variance (ANOVA) is often used to estimate relative expression of each gene in each sample (Kerr et al., 2000; Wolfinger et al., 2001).

The other approach to deal with multiple conditions is the cluster analysis (Eisen et al., 1998), which aims to classify genes with similar expression patterns into the same cluster. There are two main clustering categories: unsupervised clustering and supervised clustering also known as supervised classification. Unsupervised clustering methods do not require any underlying statistical models. The popular methods includes k-means (Tavazoie et al., 1999), support vector machines (Brown et al., 2000), self-organizing maps (Herrero et al., 2001) and so on. Unsupervised clustering refers to identification of previously unknown subgroups of samples (for example, subclasses of tumors) or groups of genes (for example, genes that are co-regulated in response to some condition). The supervised clustering methods are model-based which discriminate samples to known categories.

In traditional differential expression analysis and clustering analysis, if we need to study the relationship of gene expression and a continuous phenotype, we have to converted the phenotype into two or a few discrete ordered phenotype for further analysis, which might lose some information in this converting process. In chapter 3, we proposed a model to study the association between gene expression and a continuous phenotype without doing any partition of phenotype.

1.3.3 eQTL analysis

Traditional QTL studies has largely focused on the identification of loci affecting one, or at most a few, complex traits. Microarray technology measures thousands of gene expression which can be considered as traits in mapping studies. The study between gene expression data and multiple loci is eQTL analysis. eQTL involves a lot of statistics-related research such as experimental design, linkage study, hot spots identification, and gene network. Linkage study is the key part of eQTL study, which determines subsequent studies such as identifying hot spots, constructing gene networks, and narrowing down lists of candidate genes. The linkage study in eQTL analysis is similar as traditional QTL study in structure, but with thousands of gene expression treated as phenotype. The early linkage study in eQTL analysis made use of traditional QTL method but ignored multiplicities across transcripts. Usually a LOD profile is generated for each transcript and multiple tests are constructed for each marker. Recently, researchers have attempted to deal with the multiplicities across transcripts. A lot of efforts lie in controlling an overall FDR for single and multiple linkages. The mixture over marker(MOM) approach proposed by Kendziorski et al. (2006)is the first attempt to analyze transcripts and markers jointly. The approach considers adjustments for multiple tests across both markers and transcripts. However, the assumption that transcript is associated with

one and only one of the markers limits the application of MOM method. Jia and Xu (2007) proposed a new model and relaxed this assumption of MOM method. In chapter 3, we improve the model proposed by Jia and Xu (2007) and make it more efficient in computing time.

1.3.4 Extension of eQTL analysis to deep sequencing data

With the development of sequencing technology, the era of sequencing-based approaches is coming. These approaches are also known as second generation deep sequencing. Overall, there are three sequencing technologies developed by 454 Life Sciences (Roche) (Margulies et al., 2005), Illumina (formerly Solexa sequencing) (Bennett et al., 2005), and ABI (SOLiD sequencing) (Shendure et al., 2005). The developed technologies are widely applied in genetic variation, transcription factor binding sites, and DNA methylation. Applications to the measurement of mRNA expression levels have proceeded more slowly, partly because of difficulties in developing appropriate experimental protocols, but also because the cost of sequencing technology. However, the expression analysis based on sequencing technology is still promising comparing with former Microarray technology (Marioni et al., 2008). We think the expression analysis based on sequencing technology will be more and more popular when the cost decreases. In chapter 4, we try to step a little further. We apply eQTL analysis to deep sequencing expression data. It is impossible for us to use proposed model in eQTL analysis directly due to the structure of expression data. In Microarray technology, the generated expression data are continuous and can be considered as normal distribution. The deep sequencing expression data are the number of counts for each gene, which are considered as poisson distribution. We use pseudo data to solve the problem. Due to no real data available, we only do several simulation studies. The results of simulation studies are

pretty good. We believe our proposed method is a good start in deep sequencing eQTL study.

Chapter 2

Langmuir GCRMA model

2.1 Introduction

Since the introduction of Affymetrix's high-density system in 1996 (Lockhart et al., 1996), more than 30 statistical methods have been proposed to derive gene expression indexes from raw intensity data (Irizarry et al., 2006). Most of these methods consist of three main preprocessing steps: background correction, normalization, and summarization, except a few of them omit the background correction. Irizarry et al. (2006) benchmarked 31 algorithms using a U95A dataset of spike-in controls and found that methods that differ only in normalization result in practically identical measures. More importantly, they found that background correction has the largest effect on performance, especially for low concentration, which is the main factor explaining differences between statistical methods. Since no background correction leads to attenuated estimates of differential expression (bias), most methods have been focusing on background correction to perfect accuracy. However, background correction appears to improve accuracy but, in general, worsen precision, especially when naive background correction such as directly using MM intensity values is used . Currently, model-based probe-

specific background correction (Wu et al., 2004) has been shown to maintain the overall best accuracy with comparable precision relative to others (Wu et al., 2004; Irizarry et al., 2006).

Among all the statistical algorithms benchmarked by Irizarry et al. (2006), models used for predicting expression index (summarization step) relied on the assumption of linearity between the concentration of any target molecule and the hybridization intensity of its probe. However, non-linear hybridization behavior has been revealed by experimental results, especially when target concentration is high and/or probe affinity is strong. Moreover, it has been suggested that the fluorescence intensity measurements strongly depend on the hybridization free energy between probe and target and that the free energy can be estimated from the probe sequences (Chudin et al., 2001; Hekstra et al., 2003; Held et al., 2003b; Zhang et al., 2003; Burden et al., 2004; Abdueva et al., 2006). Probe sequences were also found to correlate with high variation of saturation levels amongst different probes (Held et al., 2006; Burden et al., 2006). Langmuir adsorption isotherm, which is based on well established principles of the physical chemistry of hybridization, has demonstrated that it can capture the nonlinear shape of GeneChip hybridization very well (Hekstra et al., 2003; Held et al., 2003b; Burden et al., 2004). Since its parameters all have physical units, a Langmuir adsorption model can also predict absolute targets concentration as opposed to gene expression indices, and hence enable the comparisons between expression levels of different genes that are forbidden by most empirical statistical models. While this line of research shows great promise, to date there have been a limited but growing number of attempts (Ono et al., 2008; Mulders et al., 2009).

In this chapter, we propose a Langmuir-type thermodynamic model which incorporates probe sequence information to predict absolute targets concentration with the

effects of saturation at high target concentration and probe sequence specificity being accounted for at the same time. For background correction, we will adapt the model-based sequence-specific background correction developed by Wu et al. (2004). Figure 3b in Irizarry et al. (2006) has shown that bias is worst for high expressed genes due to saturation. Therefore, our proposed method can reduce the biases for highly expressed genes and result in improved accuracy of absolute target concentration estimation. The resulting absolute target concentration estimation will allow not only the comparisons between different treatments of a given gene within the same experiment, but also the comparison between different genes or the same gene in different experiments.

2.2 Previous Work

It has been widely accepted that the observed raw intensity of each PM probe consists of three components: specific binding (SB), non-specific binding (NSB) and optical noise (O). The part of the observed intensity due to optical noise and nonspecific binding is usually referred to as background. Throughout the chapter, we denote the intensities obtained for each probe pair as $PM_{ij(k)}$ and $MM_{ij(k)}$, $i = 1, 2, \cdots, I$; $j = 1, 2, \cdots, J$; $k = 1, 2, \cdots, K$ with i representing the index of the RNA samples, j representing the index of the different probe sets and k representing the probe indices nested in the probe set j. Affymetrix's first attempt at an expression measure (MAS 4.0) used the following simple linear additive statistical model:

$$
\begin{aligned}
PM_{ij(k)} &= SB_{ij(k)} + BG_{ij(k)} \qquad (2.1)\\
BG_{ij(k)} &= NSB_{ij(k)} + O_{ij(k)},\\
SB_{ij(k)} &= \mu_{ij} + \varepsilon_{ij(k)},\\
i &= 1, 2, \cdots, I;\ j = 1, 2, \cdots, J;\ k = 1, 2, \cdots, K,
\end{aligned}
$$

where μ_{ij} denotes the expression index for the jth probe set hybridized with the ith RNA sample and the estimate of μ_{ij} is

$$\widehat{\mu}_{ij}^{MAS4} = \sum_{k=1}^{K} (PM_{ij(k)} - MM_{ij(k)})/K.$$

Obviously, the success of this method relies on two assumptions: that local background can be reliably estimated by the MM intensity values and that the error term $\varepsilon_{ij(k)}$ has equal variance for $k = 1, 2, \cdots, K$. However, it has been shown that probes with larger mean intensities have larger variances (Irizarry et al., 2003b). Empirical results also demonstrated MM probes detected not only local background but also partial signals, resulting in MM $\geq$ PM for about 1/3 of the probes on any given array (Naef and Magnasco; Irizarry et al., 2003b). In their second-generation algorithm MAS 5.0, Affymetrix used a log transformation on PM-CT where CT is a quantity derived from MMs that is never larger than its PM. The log of expression index μ_{ij} is then summarized by Tukey Biweight$\{log(PM_{ij(k)} - CT_{ij(k)}),\ k = 1, 2, \cdots, K\}$. Using replicate array data, Irizarry et al. (2003a) showed the decreased bias in expression indices derived by MAS 5.0 but at the price of increased variance(Irizarry et al., 2003b). Li and Wong (2001) observed strong probe affinity effects and proposed to model SB in model (2.1) as

$$SB_{ij(k)} = \mu_{ij}\phi_{j(k)} + \varepsilon_{ij(k)},$$

where $\phi_{j(k)}$ denotes the probe specific binding affinity and $\epsilon_{ij(k)}$ is assumed to be independent and normally distributed with mean 0 and common variance σ^2. μ_{ij} is then estimated by the maximum likelihood method. Their work also recognized the need for nonlinear normalization and the advantage of using multi-array summaries for detection and removal of outliers. Their findings provided guidelines for the development of many popular expression measures. However, same as MAS 4.0, this method (dChip) also suffered from the use of MM values for probe specific background correction and high mean variance dependence. Taking advantage of the findings from MAS 5.0 and Li and Wong (2001), Irizarry et al. (2003b) suggested to model SB in model (2.1) as

$$SB_{ij(k)} = \mu_{ij}\phi_{j(k)}\varepsilon_{ij(k)},$$

First, $\log(\mathrm{SB}_{ij(k)})$ was estimated by $T(PM_{ij(k)})$ which represents a background corrected, normalized and log transformed PM intensity. The background correction was based on a Exponential-Normal convolution model and normalization was through quantile normalization (Bolstad et al., 2003). Estimates of $\log(\mu_{ij})$ and $\log(\phi_{j(k)})$ are then obtained by median polish (Mosteller and Tukey, 1977). Compared against MAS 5.0 and dChip, this robust multi-array analysis (RMA) method appears to have a greater reduction in variance with a modest loss of accuracy, especially for low expression values. However, the background adjustment step in RMA ignores MM and uses a global correction. Wu et al. (2004) anticipated that a probe-specific background correction with the use of both PM and MM information might bring an extra gain in accuracy. Therefore, they proposed the use of a bivariate normal distribution for the joint distribution of PM and MM in the Exponential-Normal convolution model for background correction (GCRMA). Both PM and MM probe sequences were also incorporated in

the Exponential-Normal convolution model. Using the U95A spike-in dataset, GCRMA has been shown to have the best accuracy comparing to others including the third generation (PLIER) of Affymetrix's algorithm (Held et al., 2003a) and therefore has the best performance in differential gene expression analysis (Wu et al., 2004; Irizarry et al., 2006; McGee and Chen, 2006). However, it has also been reported that the performance of GCRMA is platform dependent (McGee and Chen, 2006).

All the methods discussed above as well as those listed in Irizarry et al. (2006) are either purely statistical or empirical and explicitly assume linearity between the concentration of any target molecule and the amount of hybridization measured by the fluorescent intensity of its probe. However, in reality linearity can only be kept within a rather narrow concentration range because of the saturation effect resulting from surface adsorption processes (Halperin et al., 2004). Irizarry et al. (2006) also noticed the inflated bias at high concentration levels among all the linear statistical models compared. Recently, there have been increasing efforts in utilizing chemical adsorption models to capture hybridization behavior of arrays, among which the Langmuir adsorption model is commonly adapted (Hekstra et al., 2003; Held et al., 2003b, 2006; Burden et al., 2004, 2006; Abdueva et al., 2006; Ono et al., 2008; Mulders et al., 2009). Langmuir adsorption theory is based on the equilibrium assumption of two competing processes driving hybridization: duplex formation between target molecules and immobilized probes (adsorption) and duplex dissociation into separate probe and target molecules (desorption). In the absence of random errors (ε), the expected SB can be described by Langmuir adsorption model as:

$$E(SB_{ij(k)}) = \alpha_{ij(k)} \frac{c_{ij}}{K_{j(k)} + c_{ij}} \tag{2.2}$$

where c_{ij} represents target concentration, $\alpha_{ij(k)}$ represents saturation intensity and $K_{j(k)}$ denotes the equilibrium constant depending on the free energy of the hybridization, the gas constant and the temperature.

Abdueva et al. (2006) considered the following statistical model:

$$PM_{ij(k)} = (\alpha_{ij(k)} \frac{c_{ij}}{K_{j(k)} + c_{ij}} + BG_{ij(k)}) e^{\epsilon_{ij(k)}} \tag{2.3}$$

and implemented background and gene expression estimation within the same fitting procedure based on a non-linear least square method. Like dChip, RMA and GCRMA, one major drawback of model (2.3) is that it requires a large number of arrays to obtain reliable estimation for all the model parameters. Held et al. (2003b) considered the same model except the saturation intensity α was assumed to be constant among all probes. However, recent studies reported a high variation of saturation levels amongst different probes which highly depends on probe sequences (Hekstra et al., 2003; Held et al., 2003b; Burden et al., 2004, 2006). Burden et al. (2004) considered the following statistical model:

$$PM_{ij(k)} \sim Gamma(\theta_{ij(k)}/\nu, \nu) \text{ with } \theta_{ij(k)} = \alpha_{ij(k)} \frac{c_{ij}}{K_{j(k)} + c_{ij}} + BG_{ij(k)} \tag{2.4}$$

where θ is the mean of the gamma distribution and ν is a constant shape parameter. They compared model (2.4) with a few extended models and concluded model (2.4) was the most parsimonious and accurate model. Hekstra et al. (2003) also considered the similar model which didn't assume any specific distribution for PM and where BG was replaced by NSB. They used the weighted least square approach to estimate parameters for the Langmuir model using known concentration values provided in a spike-in experiment. The resulting parameters were then fit into a linear combination of the

numbers of each nucleotide for each probe-target pair and estimates of concentrations were obtained for each probe by inverting the Langmuir equation. The averages of the predicted concentrations across each probeset were then reported as expression measures. Burden et al. (2004) demonstrated that such an approach returns poor estimates of concentration with up to 60% unusable predicted values. Note that for all the above methods based on Langmuir adsorption model, background (with or without optical noise corrected) and gene expression estimation were implemented with the same least square fitting procedure. Although Abdueva et al. (2006) illustrated the improved sensitivity in differential gene expression analysis using such a unified fitting procedure, the significant increase in variance in estimating concentration change was also evident.

RMA appears to have a greater reduction in variance with a modest loss of accuracy. GCRMA obtained extra accuracy and kept the same variance level as RMA with a little sacrifice in gene differential expression power. But GCRMA fails to solve the high concentration inflated bias problem. Meanwhile, a physico-chemical model such as the Langmuir model is one of the options to solve the high concentration inflated bias problem but fails to decrease the variance in estimating the concentration change. Apparently, as we have shown, these two lines of research are moving forward in a parallel manner, which therefore motivated us to incorporate a physico-chemical model similar to Burden et al. (2004) into the statistical model proposed by Wu et al. (2004) and improve the accuracy of high concentration.

2.3 Langmuir GCRMA Model

2.3.1 Proposed model

It is commonly accepted that the value of a probe intensity read from a target gene j in array i and probe k can be described by the following model (Wu and Irizarry, 2007):

$$PM_{ij(k)} = SB_{ij(k)} + NSB_{ij(k)} + O_{ij(k)} \tag{2.5}$$

with $i = 1, 2, \cdots, I$; $j = 1, 2, \cdots, J$; $k = 1, 2, \cdots, K$. $PM_{ij(k)}$ is the normalized probe signal. According to the GCRMA model, the gene specific binding component $SB_{ij(k)}$ can be further decomposed into

$$SB_{ij(k)} = \exp(c_{ij} + \phi_{j(k)} + \epsilon_{ij(k)}), \text{ if } SB_{ij(k)} > 0 \tag{2.6}$$

The gene specific binding component $SB_{ij(k)}$ is formed by a log-scale probe effect ϕ, a measurement error ε and a quantity proportional to the amount of the transcript e^c. From the model we can see that the intensity grows linearly with the amount of target if we remove the background. However, the relationship between intensity and the amount of target is nonlinear by langmuir's theory. As shown in Figure 2.1, the shape of the distribution is skewed and similar to gamma distribution, we therefore consider the following model for $SB_{ij(k)}$ as suggested by Burden et al. (2004):

$$SB_{ij(k)} \tilde{} Gamma(\theta_{ij(k)}/\nu, \nu) \text{ with } \theta_{ij(k)} = \frac{\exp(\phi_{j(k)} + c_{ij})}{\exp(c_{ij}) + \exp(d_{j(k)})} \tag{2.7}$$

Here $\theta_{ij(k)}$ is the mean of the gamma distribution. ν is a constant shape parameter. ϕ and d are probe specific binding affinities and can be obtained from the sequence

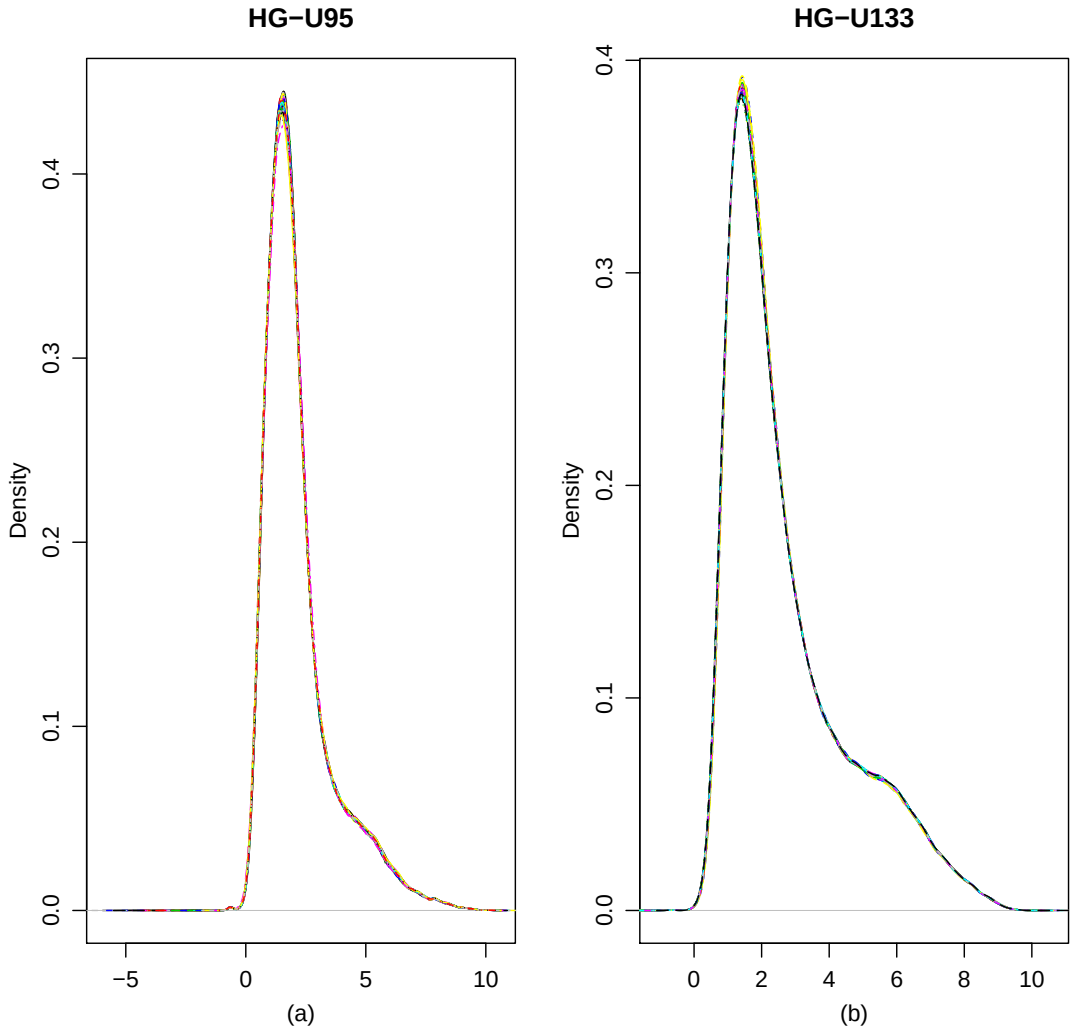

Figure 2.1: Smoothed density estimates of background corrected PM intensities from the Affymetrix Latin Square spike-in experiments. (a)The U95 data contains 59 samples (b)The U133 data contain 42 samples.

information of each probe by using a nearest neighbor (nn) model (SantaLucia, 1998).

$$\phi_{j(k)} = \sum_{m=1}^{24} w_m[\gamma_\phi(b_m, b_{m+1})] \tag{2.8}$$

$$d_{j(k)} = \sum_{m=1}^{24} w_m[\gamma_d(b_m, b_{m+1})] \tag{2.9}$$

In equation (2.8) and (2.9), we use weights w_m and base pair stacking parameters γ_ϕ and γ_d to estimate probe specific parameters ϕ and d. The probe binding affinity depends on both stacking base pair parameters and weights. First let's understand the stacking base pair parameters γ_ϕ and γ_d. From the previous introduction we know each probe will hybridize to its complimentary target when sample targets are poured to each chip and become a hybridized helix. If we treat each binding base pair of a hybridized helix as one layer, we can get 25 layers for each probe. In chemical theory, there exists stacking energy between stacking base pairs. We can get 16 different base pair stacking parameters due to 4 base types. In previous papers, Zhang et al. (2003) used all 16 base pair stacking parameters. Ono et al. (2008) used 10 parameters which divide 16 parameters into 10 groups. In order to save computing time we developed a new partition and simply got the 6 groups given by Table 2.1. Rows 2 and 3 in Table 2.1 stand for two adjacent layers of a hybridized helix. For example, in category 1, AT in row 2 is one layer of a hybridized helix and AT in row 3 is the adjacent layer. Since γ_ϕ and γ_d represent different stacking parameters, the total number of stacking parameters is 32, 20, and 12 for the three partitions mentioned above.

Table 2.1: Six groups of stacking base pairs

1	2	3	4	5	6
AT TA	AT GC TA CG	AT CG GC TA	AT TA	CG GC	CG GC
AT TA	CG TA GC AT	GC TA AT CG	TA AT	CG GC	GC CG

The next step is to understand the weights. In fact, the contribution of base pair

stacking energy is different according to different positions of a hybridized helix. The stacking base pair will contribute more if it is in the middle of the helix. So we need to use weights to measure the contributions of different positions. w_m stands for the weight between binding base pair of mth and $(m+1)$th layer and is determined by the distance between the mth and $(m+1)$th stacking base pair and the central base of the probe. Each hybridized helix consisted of 25 layers which resulted in 24 stacking base pairs and weights. Empirical studies by R.W. Michelmore's group (personal communication) suggested that the weighting factor w_m can be estimated by equation (2.10)

$$w_m = a(-0.0022x_m^4 + 0.00005x_m^3 + 0.0791x_m^2 - 0.0537x_m + 81.211) \tag{2.10}$$

where x_m for $m = 1, \cdots, 24$ denotes the relative distance between the mth and $(m+1)$th stacking base pair and the central base of a probe, ranging from -11.5 to 11.5 increased by 1. As shown in Figure 2.2, the closer a stacking base pair is to the central base of the probe, the larger the weighting factor and therefore the larger the binding contribution of the stacking base pair is. a is an unknown weight scale parameter. So every probe specific binding affinity ϕ and d is obtained by a weighted sum of the 24 base pair stacking energies.

2.3.2 Parameter estimation

Based on distribution assumption (2.7), the log-likelihood function for all probes can be written as

$$\begin{aligned} l(SB_{ij(k)}, \theta_{ij(k)}) = (\nu - 1) \sum_{i,j,k=1}^{I,J,K} \log(SB_{ij(k)}) - \sum_{i,j,k=1}^{I,J,K} \frac{\nu SB_{ij(k)}}{\theta_{ij(k)}} \\ - \nu \sum_{i,j,k=1}^{I,J,K} \log\theta_{ij(k)} - IJK(\log\Gamma(\nu) - \nu\log\nu) \end{aligned} \tag{2.11}$$

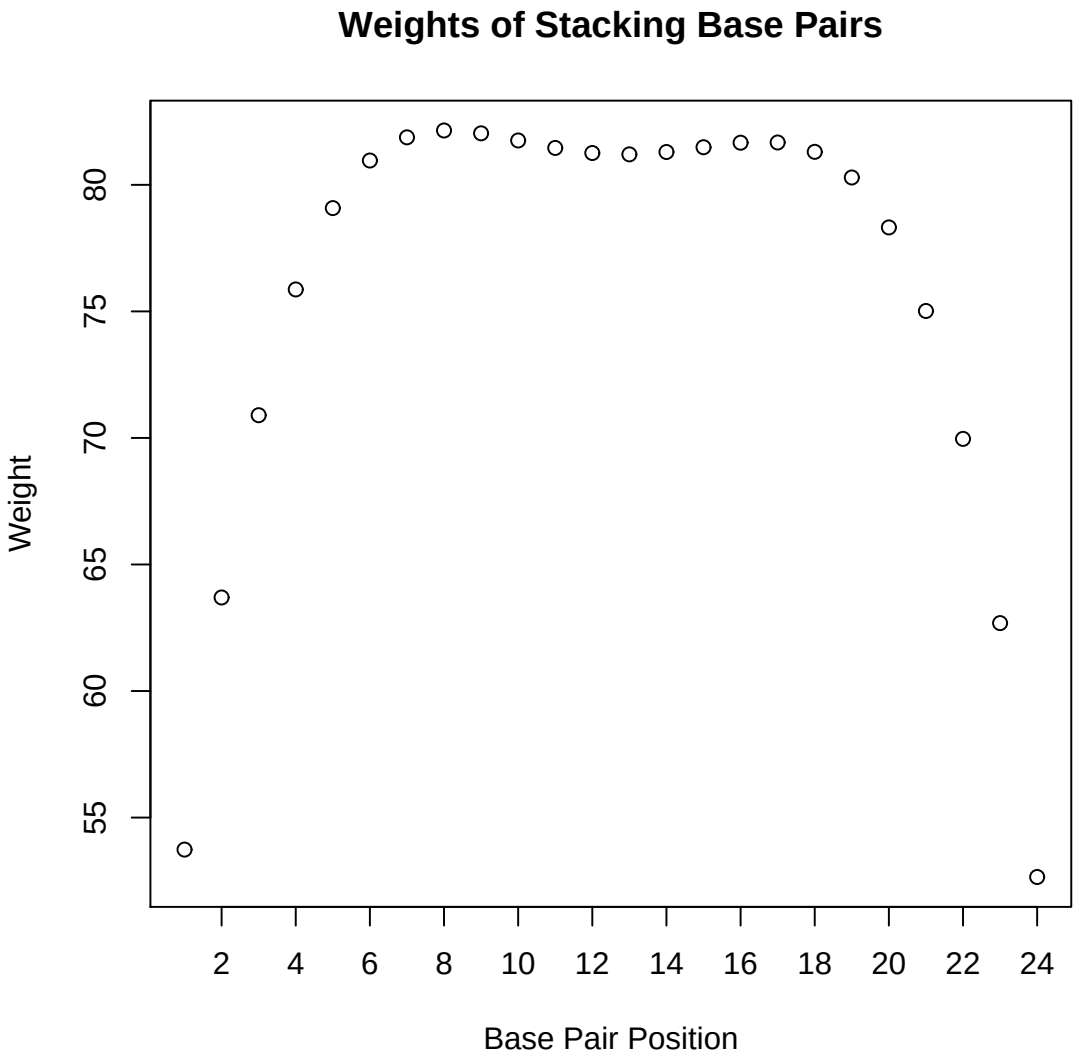

Figure 2.2: The 24 weights of the stacking base pairs.

Note that the $SB_{ij(k)}$'s in equation (2.11) are obtained by applying the GCRMA background correction to the raw intensity data. According to the formulas (2.8) ,(2.9), and table (2.1), we need to estimate 13 sequence dependent parameters – 12 stacking parameters and an unknown weight scale parameter a, unknown concentration $c_{ij}, i = 1, \cdots, I, j = 1, \cdots, J$, and shape parameter ν. Since the log-likelihood function is highly complex and no closed form is available for calculating the MLEs of the parameters, a numerical optimization method has to be applied to find the MLEs. Furthermore, there are a large number of parameters that must be estimated. If we consider a toy example, say 100 genes and 10 samples, the parameter space is still very large. We have 13 sequence dependent parameters, one shape parameter, and $100 \times 10 = 1000$ unknown concentrations. So the total number of parameters is $13+1+1000 = 1014$. Many optimization algorithms frequently suffer from slow convergence due to the search over the large parameter space. Welinan (1993) suggested to break the original parameter space into subspaces and perform an optimization algorithm on each subspace, one at a time in a iterative manner. Therefore, we proposed the following iterative estimation procedure:

1. Assign initial value ν and c_{ij}

2. Update γ_ϕ, γ_d and a given ν and c_{ij} using the function 'optim()', a multiple dimensional parameter estimation routine in R:

$$\text{minimize} \sum_{i,j,k=1}^{I,J,K} \left[\frac{SB_{ij(k)}}{\theta_{ij(k)}} - \log \frac{SB_{ij(k)}}{\theta_{ij(k)}} \right] \xrightarrow{\text{given } \nu, c_{ij}} \gamma_\phi, \gamma_d, a$$

3. Update each c_{ij} given all the other parameters using the function 'optimize()', a one-dimensional parameter estimation routine:

$$\text{minimize} \sum_{k=1}^{K} \left[\frac{SB_{ij(k)}}{\theta_{ij(k)}} - \log \frac{SB_{ij(k)}}{\theta_{ij(k)}} \right] \xrightarrow{\text{given } \gamma_\phi,\ \gamma_d,\ \nu} c_{ij}$$

4. Update ν given all other parameters using the function 'optimize()', a one-dimensional parameter estimation routine:

$$\text{minimize} \sum_{i,j,k=1}^{I,J,K} \nu \left[\frac{SB_{ij(k)}}{\theta_{ij(k)}} - \log \frac{SB_{ij(k)}}{\theta_{ij(k)}} \right] + IJK \left(\log\Gamma(\nu) - \nu \log\nu \right) \xrightarrow{\text{given } \gamma_\phi, \gamma_d, c_{ij}} \nu$$

5. Repeat steps 1-4 and stop until the log-likelihood l converges. We stop the iteration when the difference of concectutive l values is less than 10^{-5}.

We selected some genes to do the above procedure due to the computing intensity of the huge data set. The selection criteria will be discussed in the spike-in data analysis section. When we get the final parameters γ_ϕ, γ_d, and ν from the selected genes, we will apply these parameters in step 3 to estimate the unselected gene concentration c_{ij}.

2.4 Results

2.4.1 Simulation study

We did a simulation study to assess the performance of the above estimation procedure. 91 genes were selected from the U95 data. The concentration $c_{ij}, i = 1, \cdots, 10, j = 1, \cdots, 91$ is evenly sampled from 0 to 12 (log2 scale), which means 7 genes share the same concentration. A gene from different samples share the same concentration as well. Sequence dependent parameters γ_ϕ, γ_d are given in Table 2.2 and the weight scale a equals to 0.01. The constant shape parameter ν is 10. The raw data $SB_{ij(k)}$ can be generated based on the above given parameters. We replicate the above simulation 20 times and estimate the concentration c_{ij} by our proposed model L-GCRMA and RMA. Since our motivation is to improve accuracy at high concentration, we expect the slope of our proposed L-GCRMA model to be more accurate than the slope of the RMA model at high concentration. Figure 2.3 shows the average estimated versus nominal concentrations of L-GCRMA and RMA across 20 replicates. The slope of L-GCRMA

Table 2.2: Sequence dependent parameters γ_ϕ, γ_d in simulation

γ_ϕ	0.54	0.55	0.48	0.68	0.06	0.61
γ_d	0.43	0.45	0.46	0.63	-0.11	0.51

matches perfectly with the identity line. While the slope of RMA decreases at high concentration. Therefore the simulation study demonstrate the advantage of L-GCRMA at high concentration.

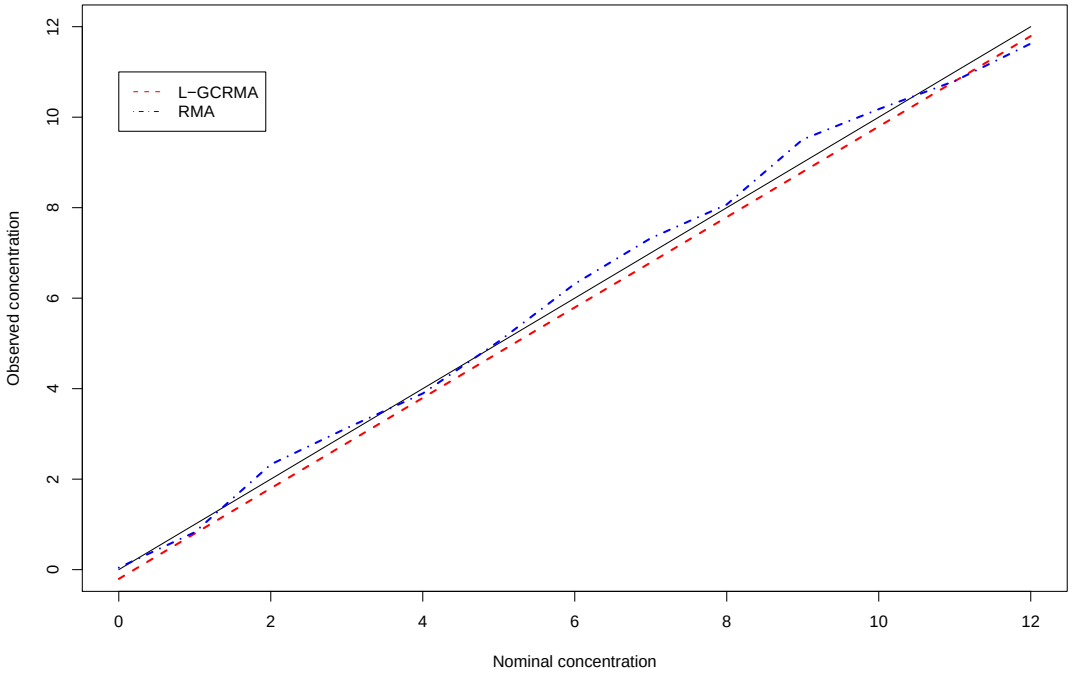

Figure 2.3: Estimated concentration versus nominal concentration in simulation

2.4.2 Spikein data analysis

We used two Affymetrix spike-in data sets to test the performance of our proposed model.

Affymetrix Human genome U95 dataset. This dataset contains 59 arrays organized in a Latin square design. 14 groups of human genes are spiked in at a known concen-

tration. The concentrations of the 14 gene groups are 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024 pM. Each subsequent array group rotates the spike-in concentrations by one group, i.e. array group 2 begins with 0.25 pM and ends at 0 pM, and this pattern continues until array group 14, which begins with 1024 pM and ends with 512 pM. Each experiment has three replicates except for experiment 13 and 14 which consist of 12 replicates. Meanwhile, one replicate of experiment 3 is missing. So in total there are 59 arrays. The number of spike-in genes in this data is 14. In addition, other researchers have found that probe set 546_at and 33818_at should also be considered as spike-in genes. Therefore in our analysis, we included these 2 genes, resulting in a total of 16 spike-in genes. Table 2.3 shows the original experiment design of the U95 spike-in data.

Table 2.3: U95 spike-in experiment

Gene ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10	Exp11	Exp12	Exp13	Exp14
37777_at	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
684_at	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024	0
1597_at	0.5	1	2	4	8	16	32	64	128	256	512	1024	0	0.25
38734_at	1	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5
39058_at	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1
36311_at	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2
36889_at	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4
1024_at	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8
36202_at	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16
36085_at	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32
40322_at	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
407_at	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
1091_at	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
1708_at	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

Affymetrix Human genome U133 dataset. The structure of the U133 data is similar to the U95 dataset. This dataset consists of 3 technical replicates of 14 experiments

which contain 42 spike-in genes. The concentrations of the 42 spike-in genes are ranging from 0 to 512 pM. Since 22 additional spike-in genes were found by other researchers, we include these 22 additional spike-in genes in our analysis. Table 2.4 shows the original experiment design of the U133 spike-in data.

Table 2.4: U133 spike-in experiment

Gene ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10	Exp11	Exp12	Exp13	Exp14
203508_at 204563_at 204513_s_at	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512
204205_at 204959_at 207655_s_at	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0
204836_at 205291_at 209795_at	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125
207777_s_at 204912_at 205569_at	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25
207160_at 205692_s_at 212827_at	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5
209606_at 205267_at 204417_at	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1
205398_s_at 209734_at 209354_at	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2
206060_s_at 205790_at 200665_s_at	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4
207641_at 207540_s_at 204430_s_at	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8
203471_s_at 204951_at 207968_s_at	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16
AFFX-r2-TagA_at AFFX-r2-TagB_at AFFX-r2-TagC_at	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32
AFFX-r2-TagD_at AFFX-r2-TagE_at AFFX-r2-TagF_at	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64
AFFX-r2-TagG_at AFFX-r2-TagH_at AFFX-DapX-3_at	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128
AFFX-LysX-3_at AFFX-PheX-3_at AFFX-ThrX-3_at	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256

The criterion to select which genes to estimate parameters γ_ϕ, γ_d and ν is based on gene variance. Since our data $SB_{ij(k)}$ is probe level data, the definition of gene variance is the median of probe variance of that gene. For example, we can first calculate the

probe variance across 42 samples in U133. Then the gene variance is the median of its 11 probe variances. Genes with large variance will be selected to estimate parameters. This criterion can give us more information to estimate the sequence dependent parameters γ_ϕ, γ_d for the same number of selected genes. Let's consider a simple case. Suppose we only select one gene including 11 probes and 10 samples to do the estimation. If the selected gene variance is very small, then we actually only use 11 data points to estimate the parameters γ_ϕ, γ_d and ν due to small differences among samples for each probe. If the selected gene variance is large, we can use $11 \times 10 = 110$ data points to estimate the parameters. The number of selected genes depends on the data used. In our analysis, we selected 1000 genes to estimate the parameters γ_ϕ, γ_d and ν.

To assess the accuracy of our method, nominal against the estimated concentrations of the spike-in genes for the HU133 and HU95 data were plotted in Figure 2.4a and 2.4c respectively, in which the red lines represent the average value of spike-in genes and fit identity lines very well. To better assess the concentration dependent bias, we also calculated the local slopes by taking the difference between the average observed log expression values between consecutive nominal concentration levels. The differences between 1 and these local slopes are plotted against the larger of the two concentration levels in Figure 2.4b and 2.4d for the HU133 and HU95 data respectively. As can be seen, our method still demonstrate bias for both low and highly expressed genes. We further compare our method with four other popular preprocessing methods: RMA (Irizarry et al., 2003b), GCRMA (Wu et al., 2004), Plier (Held et al., 2003a), and DFCM (Zhongxue et al.). Since there are two versions of background correction in GCRMA called the adhoc and ebayes methods, we denote our methods as Langmuir-GCRMA adhoc and Langmuir-GCRMA ebayes respectively. Since the langmuir model accounts for the saturation effect at high concentration and GCRMA is considered as

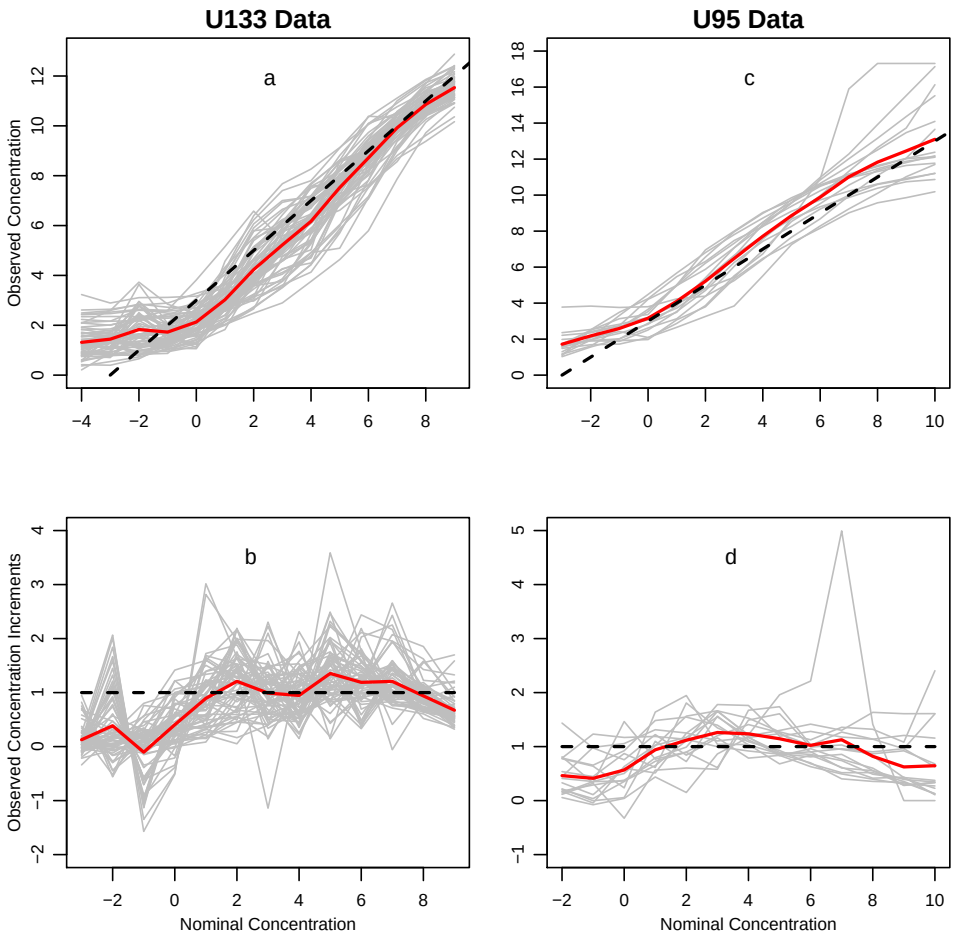

Figure 2.4: **(a),(c)**Estimated log expression versus nominal concentration of spike-in genes in U133 and U95. **(b),(d)** Local slopes versus nominal concentration of spike-in genes in U133 and U95. Red lines denote the average of the spike-in genes. Dashed lines represent identity lines.

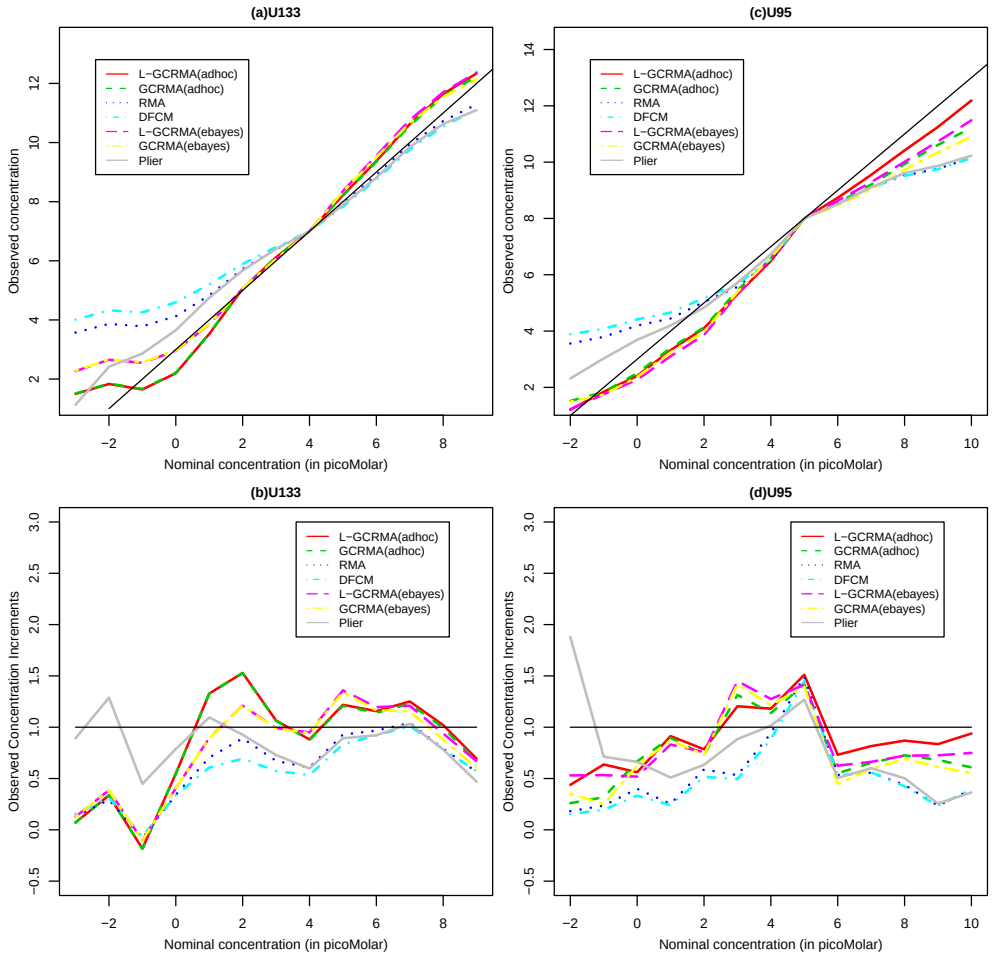

Figure 2.5: **(a),(c)**Comparison of estimated log expression versus nominal concentration of spike-in genes. **(b),(d)** Comparison of local slopes versus nominal concentration of spike-in genes. The sequence dependent parameters were estimated based on 1000 genes with our method.

the best background correction method for low concentration, we expect our method will improve the signal detection slope in Figure 2.5 at both high and low concentration level. Therefore, we stratified the spiked-in genes into low expressed (nominal concentration $<$ 4 pM), medium expressed (nominal concentration between 4 and 32) and high expressed (nominal concentration $>$ 32) and for each of these subgroups we followed the same procedure used in Irizarry et al. (2006) to compute the signal detect slope. The slopes obtained are referred to as low, median and high slopes and are shown in Table 2.5. At low and median concentrations, our proposed methods keep almost the same slope

as the GCRMA model and outperform than other three methods except Plier at low concentration. At high concentration, our methods increase GCRMA's slope, especially in the U95 data since its highest concentration is twice as much as U133's. Overall, our proposed method performs more accurately in high concentrations than other methods. It is also evident in Figure 2.5 that our Langmuir-GCRMA method performs the best in reducing the concentration dependent bias for low and high expressed genes.

Table 2.5: Slope comparison in seven methods for U95A and U133

Method	U95			U133		
	Low	Median	High	Low	Median	High
RMA	0.349	0.762	0.473	0.277	0.723	0.798
GCRMA(adhoc)	0.689	1.065	0.559	0.441	1.031	0.967
L-GCRMA(adhoc)	0.693	1.113	0.755	0.442	1.037	0.992
GCRMA(ebayes)	0.598	1.136	0.501	0.346	1.076	0.879
L-GCRMA(ebayes)	0.599	1.179	0.721	0.346	1.085	0.939
DFCM	0.297	0.728	0.468	0.265	0.635	0.788
Plier	0.723	0.838	0.465	0.848	0.725	0.764

Affymetrix GeneChip array is mainly used to detect differentially expressed genes under different experimental conditions. To compare the overall detection ability of our Langmuir-GCRMA method to others, we obtained the Receiver Operator Characteristic (ROC) curves based on the fold change filtering rule as suggested by Cope et al. (2004). Since only 16 and 64 spiked-in genes are actually differentially expressed in the HU95 and HU133 data respectively, it is easy to determine true positives (TP) and false positives (FP). Specifically, for each pair of arrays in which the nominal fold change for spike-in genes equals to 2, we ordered the probesets by the observed absolute value of their log ratios and counted the number of TPs for every possible value of $1, 2, \cdots, 100$ FPs (100 non-spiked in probe sets). For example, in the HU133 data, 14 experiments

were replicated three times and concentrations of most spike-in genes in two adjacent experiments were differentiated two fold, except for a few spike-in genes with zero concentration in one experiment and the largest concentration in the other experiment. Therefore, there are 42 pairs of arrays and we can obtain TPs for each pair of arrays as described above. For each FP value, TPs were then averaged across all 42 pairs of arrays and an average ROC curve was created (Figure 2.6a) in which the proportion of average TPs over all 64 known TPs were plotted against the FPs. Figure 2.6b is the average ROC curve for the HU95 data. Note that here we only consider the maximum of 100 false positives because lists of genes with more errors are not typically useful (Irizarry et al., 2006). Our proposed method improves GCRMA gene detection ability according to average ROC curves in the U133 and U95 data. The improvement in the U95 is obvious and makes L-GCRMA the best method. While the improvement in the U133 is not as good as that in the U95 and makes L-GCRMA comparable with DFCM. However, since the motivation of our proposed model is to improve the accuracy at high concentrations, we are more concerned about ROC curves at high concentrations. We did the same procedure as mentioned by Irizarry et al. (2006) to obtain ROC curves at high concentrations in Figure 2.6c,d for the U133 and U95 data (We only do the adhoc version since it is better than ebayes version in both GCRMA and L-GCRMA). Notice L-GCRMA improves GCRMA ROC curve dramaticlly and can be considered as the best differentially expressed gene detection method in the U133 and U95 at high concentration.

2.4.3 Arabidopsis data analysis

We applied our method to detect differentilly expressed genes in Arabidopisis data from He et al. (2009) in which the RNA-directed DNA methylation mechanism in Ara-

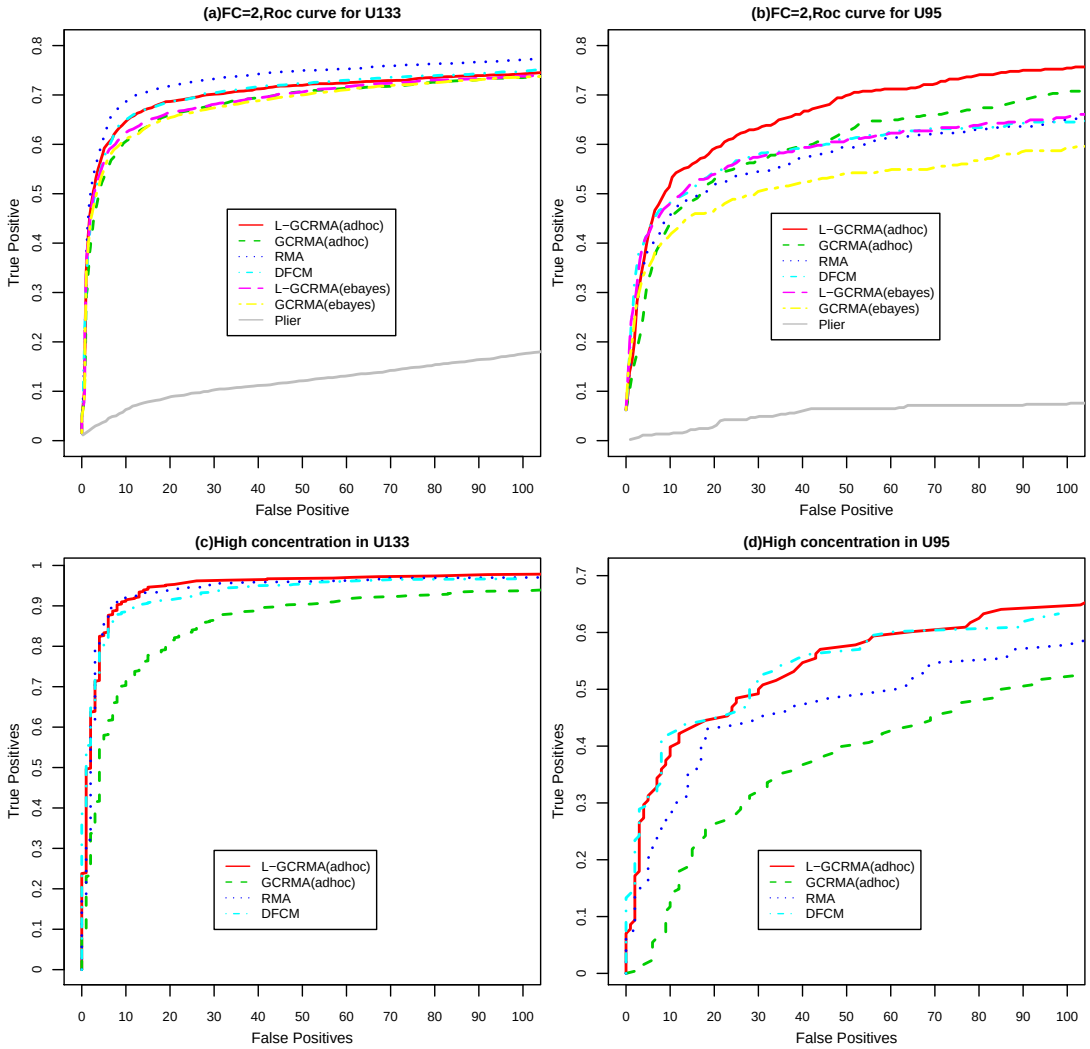

Figure 2.6: **(a)**Average ROC curves with 2 fold change in the U133 data. **(b)** Average ROC curves with 2 fold change in the U95 data. **(c)**ROC curves at high concentration with 2 fold change in the U133 data. **(d)** ROC curves at high concentration with 2 fold change in the U95 data.

bidopsis was studied. In this study, the Affymetrix Arabidopsis ATH1 GeneChip was used. There are two genotypes, ros1 and ros1rdm4, each of which consists of two biological replicates. In their paper, they used the RMA preprocessing method and SAM to detect differentially expressed genes. Meanwhile, real-time PCR and northern blotting analysis were also performed to validate four differentially expressed genes (ROS1, COR15A, P5CS and KIN1,2) which were identified by the statistical analysis. We used our proposed method and the GCRMA method for preprocessing and combining with SAM to detect differentially expressed genes. Table 2.6 shows the fold change for four genes obtained from the five methods mentioned above. As can be seen, the order of the four genes ranked by fold change using our proposed method matches perfectly with the order obtained from real-time PCR and northern blotting analysis. On the other hand, the fold change of the gene COR15A is the smallest in the RMA and GCRMA methods. However, the results from the real-time PCR and northern blotting shows that its fold change is larger than that of P5CS or KIN1,2. The FDR of our proposed method is relatively smaller than that of RMA or GCRMA.

Table 2.6: Fold change obtained from five methods.

Gene ID	L-GCRMA	RMA	GCRMA	Northern blotting	PCR
263909_at	Fold=4.2	Fold=3.8	Fold=9.1	-	Fold=7.4
(ROS1)	FDR=0.136	FDR=0.135	FDR=0.158	-	-
263497_at	Fold=2.3	Fold=2.1	Fold=2.6	Fold=4.6	Fold=3.3
(COR15A)	FDR=0.144	FDR=0.161	FDR=0.160	-	-
251775_s_at	Fold=2.1	Fold=2.5	Fold=3.0	Fold=2.0	-
(P5CS)	FDR=0.145	FDR=0.157	FDR=0.162	-	-
246481_s_at	Fold=2.0	Fold=2.3	Fold=2.6	Fold=1.7	Fold=2.3
(KIN1,2)	FDR=0.151	FDR=0.161	FDR=0.163	-	-

2.5 Discussion

In this chapter, we proposed a Langmuir-GCRMA model to estimate gene expressions for microarray data. We first used GCRMA model to do the background correction. Then the Langmuir model was applied to estimate the gene expressions. This model combined the advantage of the GCRMA model and the Langmuir model. The GCRMA model can be considered one of the best background correction methods and can improve the accuracy of gene expression estimation. However, after the background correction, it estimates gene expressions with a linear assumption not proved in many experiments. The linear assumption will result in inaccurate gene expression at high concentration. Therefore we considered the Langmuir model which is built on a nonlinear assumption and it improved the accuracy at high concentration. The comparison of our proposed method with GCRMA demonstrates the advantage of our model at high concentration both in accuracy and in differential gene expression analysis. Overall, our model can be considered as the best model keeping the balance of accuracy and precision.

In our proposed model, we estimate binding affinity for each probe with sequence dependent parameters. In SFP detection binding affinity is a big issue. It is necessary to distinguish between a low-intensity signal due to poor hybridization resulting from a sequence polymorphism and a low-intensity signal due to low gene expression in the SFP detection problem. Ronald et al. (2005) solved the above problem by comparing the observed probe signal with the expected probe signal obtained from the PDNN model. We believe that our proposed model can also be applied in SFP detection to solve a similar problem.

Chapter 3

SEM algorithm for Microarray Study

3.1 introduction

The recently developed microarray technology allows us to measure the expression of many genes or transcripts in a single chip. Mendelian loci in the genome that control the expression levels of transcripts are called expression quantitative trait loci (eQTL). In eQTL studies, a linkage from a gene expression trait to a locus is referred as cis - linkage if the locus is close to the gene itself. Otherwise it is referred as trans - linkage. The purpose of a linkage study is to identify the cis - and trans - linkages between transcripts and loci. Results from the eQTL study may provide more detailed information about the biological processes of the gene network than the classical QTL study.

In early eQTL study, the eQTL mapping has been treated as either a QTL mapping problem (Lander and Botstein, 1989; Zeng, 1994; Kao et al., 1999) for multiple traits or a microarray differential expression problem (Pan, 2002; Newton et al., 2004) for multiple treatment comparisons. The mixture over marker (MOM) approach developed

by Kendziorski et al. (2006) is the first attempt to analyze transcripts and markers jointly. However, MOM approach assumed a transcript is either associated with one and only one marker or not associated with any markers at all, which means the approach can detect either the cis-locus or one of the trans-loci, but not both. Jia and Xu (2007) believed the assumption was too stringent and proposed Ebayes method. The Ebayes method is a Bayesian clustering method that analyzes all expressed transcripts and markers jointly in a single model. The big contribution of this method is that a transcript may be simultaneously associated with multiple markers and meanwhile a marker may simultaneously alter the expression of multiple transcripts. They use Markov Chain Monte Carlo (MCMC) to estimate each variable. However, MCMC sampling is very time consuming and needs huge computer intensity. In this chapter, we proposed Stochastic Expectation Maximization (SEM) algorithm (Celeux and Diebolt, 1985). In stochastic version of the EM algorithm, a stochastic algorithm is used to perform the necessary approximations in the E-step. A major limitation of the EM algorithm (DEMPSTER et al. 1977) is that whilst convergence to a stationary point of the likelihood function can be shown, this is not necessarily the global maximum. The motivation for the stochastic EM algorithm is to overcome this limitation. In our study, we need to decide whether a transcript is associated with a maker or not, which will involve an indicator variable. The traditional EM algorithm is hard to estimate the indicator variable since it might converge to a stationary point. SEM algorithm will solve this problem successfully. Meanwhile the SEM algorithm will dramatically reduce the computer burden since it does not require a lot of iterations to estimate parameters.

Zhan et al. (2010) applied the idea from Ebayes method to study the association between transcripts and a continuous phenotype. They aim to develop a new statistical method to cluster expressed genes based on their association with a quantitative trait

phenotype. The model is different from a differential expression analysis (Kerr et al., 2000; Wolfinger et al., 2001; Cui and Churchill, 2003) which can only be applied to binary phenotype to detect genes. In Zhan's paper phenotypic value were adjusted between -1 and 1 which is comparable to a marker in Ebayes method. So the proposed model by Zhan et al. (2010) is a simplified single marker Ebayes method. They applied their model to a real dataset collected in the North American Barley Genome Project. However, the result of their model showed plenty of associations between differential expressed transcripts and the phenotype. These transcripts have obvious two clusters which may not show the linear correlation with continuous phenotype since the phenotype is not clustered. We proposed adjusted SEM model accounting for two intercepts for each transcript. Each subject within a transcript would be assigned to either one of the intercepts. The results of our method reduced the number of detected transcripts with two clusters.

In this chapter, we will first discuss SEM model for eQTL study and modify the model to make it suitable for association study between transcripts and phenotype. Three simulation studies will be performed to compare our proposed model with other existing models. Finally we will apply the proposed model to analyze real barley data.

3.2 Theory and Method

3.2.1 Multiple eQTL model

Let M be the number of transcripts and N be the number of subjects(individuals) in microarray experiment. Define $y_j = [y_{j1}, \cdots, y_{jN}]^T$ for $j = 1, \cdots, M$ as an $N \times 1$ vector for transcript jth gene across N individuals. Let $Z_k = [Z_{k1}, \cdots, Z_{kN}]^T$ be an $N \times 1$ vector for the genotype indicator variables for marker k, $\forall k = 1, \cdots, p$,

where p is the total number of markers included in the model. The genotype indicator variable for individual i is defined as $Z_{ki} = \{-1, 1\}$ for the two genotypes of a backcross (BC) individual or $Z_{ki} = \{-1, 0, 1\}$ for the three genotypes of an F_2 individual. The expressions of gene j from all N individuals, y_j, is described by the following linear model,

$$y_j = 1\beta_j + \sum_{k=1}^{p} Z_k \gamma_{jk} + \varepsilon_j \tag{3.1}$$

where 1 is an $N \times 1$ vector of unity, β_j is the intercept(a scalar), γ_{jk} (a scalar) is the eQTL effect of transcript j for the marker k and $\varepsilon_j = [\varepsilon_{j1}, \cdots, \varepsilon_{jN}]^T$ is an $N \times 1$ vector for the residual errors with an assumed multivariate $N(0, I\sigma^2)$ distribution. Let us assign a normal distribution to β_j so that $p(\beta_j) = N(\beta_j|\mu_\beta, \sigma^2_\beta)$, where μ_β and σ^2_β are the unknown mean and variance of β_j. In this study we assign a Gaussian mixture to γ_{jk} so that

$$p(\gamma_{jk}) = (1 - \pi_k)N(\gamma_{jk}|0, \sigma^2_0) + \pi_k N(\gamma_{jk}|0, \sigma^2_k)$$

where π_k is the proportion of transcripts that belong to cluster one for locus k (the cluster for the associated transcripts), σ^2_k is an unknown variance assigned for γ_k across all transcripts, and $\sigma^2_0 = 10^{-10}$is a small positive number representing the neutral cluster (the cluster for transcripts not associated with marker k). Note there are two clusters for each locus, cluster zero, indicated by σ^2_0, and cluster one, indicated by σ^2_k. All transcripts that are classified into cluster one are associated with marker k. Since there are two clusters for each locus, we introduce an indicator variable, η_{jk}, to represent the class label, which has a Bernoulli prior distribution, i.e., $p(\eta_{jk}) = \text{Bernoulli}(\eta_{jk}|\pi_k)$. The variance of genes classified into cluster one is assigned a scaled inverse chi-square distribution, denoted by $p(\sigma^2_k) = \text{Inv} - \chi^2(\sigma^2_k|d_0, \omega_0)$, where d_0 and ω_0 are the prior degree of freedom and prior scale parameter. The proportion of associated transcripts

to marker k is also treated as a parameter with a flat prior $p(\pi_k) = \text{Beta}(0, 0)$.

The purpose of this analysis is to estimate π_k, $\forall k = 1, \cdots, p$ and the posterior mean of η_{jk}, $\forall j = 1, \cdots, M \& k = 1, \cdots, p$. The estimated π_k indicates whether or not locus k is a hot spot, whereas the estimated η_{jk} indicates whether or not transcript j is associated with locus k.

3.2.2 Gene-trait association model

This model is to identify the association of transcripts (called genes in this study) and the phenotype of a quantitative trait, and thus the model is referred to as the gene-trait association model. The gene expression vector is still denoted by y_j for gene j, but the independent variable $Z = [Z_1, \cdots, Z_N]^T$ is a vector for the rescaled phenotypic values of a quantitative trait for the N individuals. The rescaling is conducted through the following equation,

$$Z_i = 2\frac{Z_i^* - Z_{\min}^*}{Z_{\max}^* - Z_{\min}^*} - 1$$

where Z_i^* is the original phenotypic value for the ith subject, $Z_{\min}^*$ and $Z_{\max}^*$ are the minimum and maximum values of the phenotypes, respectively. The rescaled Z ranges from -1 to 1, similar to the scale of genotypic indicator variables in the multiple eQTL model. The gene-trait association model is now defined as

$$y_j = X_j\beta_j + Z\gamma_j + \varepsilon_j \tag{3.2}$$

where X_j is an $N \times 2$ unknown design matrix (to be described later), $\beta_j = [\beta_{j1}, \beta_{j2}]^T$ is a 2×1 vector of intercepts, γ_j is the regression coefficient of the expression of gene j on the phenotype, and ε_j is an $N \times 1$ vector for the residual errors with an assumed $N(0, I\sigma^2)$ distribution.

We now describe the distribution for each model effect. The association between

gene j and the trait is represented by γ_j (a scalar), which is assumed to be of Gaussian mixture,

$$p(\gamma_j) = (1-\pi)N(\gamma_j|0,\sigma_0^2) + \pi N(\gamma_j|0,\sigma_1^2)$$

where π is an unknown proportion of genes associated with the trait, $\sigma_0^2 = 10^{-10}$ remains a small positive number and σ_1^2 is an unknown variance for all genes associated with the trait. The gene class label is denoted by η_j with a Bernoulli distribution $p(\eta_j) = \text{Bernoulli}(\eta_j|\pi)$. The priors for π and σ_1^2 remain the same as the ones described in the multiple eQTL model. The intercept β_j is now defined as a 2×1 vector with a multivariate normal distribution, $p(\beta_j) = N_2(\beta_j|\mu_\beta, \Sigma_\beta)$, where

$$\mu_\beta = \begin{bmatrix} \mu_1 \\ \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma_\beta = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$$

The reason for using two intercepts to describe the gene expression is based on our past experience of microarray data analysis, where some genes are often expressed in two drastically different levels. Such genes are better modeled with two intercepts. The design matrix X_j is an $N \times 2$ unknown matrix partitioned into $X_j = [X_j^{(1)}, X_j^{(2)}]$. Each element of the matrix is defined as a binary indicator variables,

$$X_{ji}^{(1)} = \begin{cases} 1 & \text{if} \beta_{j1} \text{is the intercept} \\ \\ 0 & \text{if} \beta_{j2} \text{is the intercept} \end{cases}$$

and $X_{ji}^{(2)} = 1 - X_{ji}^{(1)}$, equivalent to $X_{ji}^{(1)} + X_{ji}^{(2)} = 1$, for $i = 1, \cdots, N$. This constraint implies that $X_{ji}^{(1)}$ is a Bernoulli variable and can be modeled by $p(X_{ji}^{(1)}) = \text{Bernoulli}(X_{ji}^{(1)}|\phi_j)$, where ϕ_j is the proportion of individuals who should be modeled with intercept β_{j1} for gene j for $j = 1, \cdots, M$. This gene-specific proportion ϕ_j is a nuisance parameter described by a flat prior $p(\phi_j) = \text{Beta}(0,0)$, $\forall j = 1, \cdots, M$.

The parameters of interest in the gene-trait association study are π and the posterior mean of η_j, $\forall j = 1, \cdots, M$, where π represents the proportion of genes associated with the trait and η_j indicates the strength of the association between gene j with the phenotype of the trait.

The major differences between the eQTL model and the gene-trait association model are: (1) the intercepts of the models are different; (2) for the association part, the gene-trait association model is equivalent to a single eQTL model with $p = 1$ and the Z_k (genotype indicator variable) is replaced by Z (the phenotypic value of the trait).

3.2.3 Expectation Maximization (EM) algorithm

3.2.3.1 Multiple eQTL

The EM algorithm requires a clear distinguish between parameters and missing values. The parameter array is denoted by

$$\theta = \{\mu_\beta, \sigma^2_\beta, \sigma^2_1, \cdots, \sigma^2_p, \pi_1, \cdots, \pi_p, \sigma^2\}$$

The missing values are

$$\xi = \{\beta_j, \gamma_{jk}\}, \forall j = 1, \cdots, M \& k = 1, \cdots, p$$

If η_{jk} is known for $j = 1, \cdots, M$ and $k = 1, \cdots, p$, the multiple eQTL model is a typical mixed model problem. An EM algorithm for mixed model is already available and we can adopt it to the eQTL study. Ignoring the derivation, we simply present the EM steps here.

Step 0: Set $t = 0$ and initialize all parameters $\theta = \theta^{(t)}$.

Step 1: Calculate the posterior mean and posterior variance for β_j, $\forall j = 1, \cdots, M$, using

$$\hat{\beta}_j = E(\beta_j | \cdots) = \mu_\beta + \sigma^2_\beta 1^T V_j^{-1}(y_j - 1\mu_\beta)$$

and

$$\hat{W}_j = \text{var}(\beta_j|\cdots) = \sigma_\beta^2 - \sigma_\beta^4 1^T V_j^{-1} 1$$

where

$$V_j = \text{var}(y_j|\eta_j) = 1\sigma_\beta^2 1^T + \sum_{k=1}^{p} Z_k[\eta_{jk}\sigma_k^2 + (1-\eta_{jk})\sigma_0^2]Z_k^T + I\sigma^2$$

Step 2: Calculate the posterior mean and posterior variance for γ_{jk}, $\forall j = 1, \cdots, M \& k = 1, \cdots, p$, using

$$\hat{\gamma}_{jk} = E(\gamma_{jk}|\cdots) = \Theta_{jk} Z_k^T V_j^{-1}(y_j - 1\mu_\beta)$$

and

$$\hat{S}_{jk} = \text{var}(\gamma_{jk}|\cdots) = \Theta_{jk} - \Theta_{jk} Z_k^T V_j^{-1} Z_k \Theta_{jk}$$

where

$$\Theta_{jk} = \text{var}(\gamma_{jk}|\eta_{jk}) = \eta_{jk}\sigma_k^2 + (1-\eta_{jk})\sigma_0^2$$

Step 3: Update μ_β using

$$\mu_\beta = \left[\sum_{j=1}^{M} 1^T V_j^{-1} 1\right]^{-1} \left[\sum_{j=1}^{M} 1^T V_j^{-1} y_j\right]$$

Step 4: Update σ_β^2 using

$$\begin{aligned} \sigma_\beta^2 &= \frac{1}{M}\sum_{j=1}^{M} E\left[(\beta_j - \mu_\beta)^T(\beta_j - \mu_\beta)\right] \\ &= \frac{1}{M}\sum_{j=1}^{M} \left[(\hat{\beta}_j - \mu_\beta)^T(\hat{\beta}_j - \mu_\beta) + \hat{W}_j\right] \end{aligned}$$

Step 5: Update σ_k^2 for $k = 1, \cdots, p$ using

$$\begin{aligned} \sigma_k^2 &= \frac{1}{\pi_k M + d_0}\left[\sum_{j=1}^{M} \eta_{jk} E(\gamma_{jk}^2) + \omega_0\right] \\ &= \frac{1}{\pi_k M + d_0}\left[\sum_{j=1}^{M} \eta_{jk}(\hat{\gamma}_{jk}^2 + \hat{S}_{jk}) + \omega_0\right] \end{aligned}$$

Step 6: Update σ^2 using

$$\begin{aligned}\sigma^2 &= \frac{1}{NM}\sum_{j=1}^{M} E\left[(y_j - 1\beta_j - \sum_{k=1}^{p} Z_k\gamma_{jk})^T(y_j - 1\beta_j - \sum_{k=1}^{p} Z_k\gamma_{jk})\right] \\ &= \frac{1}{NM}\sum_{j=1}^{M}\left[(y_j - 1\hat{\beta}_j - \sum_{k=1}^{p} Z_k\hat{\gamma}_{jk})^T(y_j - 1\hat{\beta}_j - \sum_{k=1}^{p} Z_k\hat{\gamma}_{jk}) + \sum_{k=1}^{p} Z_k^T Z_k \hat{S_{jk}} + N\hat{W}_j\right]\end{aligned}$$

Step 7: Increment t by one and repeat from Step 1 to Step 6 until a certain criterion of convergence is reached.

Note that Steps 1-2 represent the E-steps and Steps 3-6 represent the M-steps, explaining why the algorithm is called EM.

3.2.3.2 Gene-trait association

The parameter vector is

$$\theta = \{\mu_\beta, \Sigma_\beta, \sigma_1^2, \pi, \sigma^2\}$$

The missing values are

$$\xi = \{\beta_j, \gamma_j\}, \forall j = 1, \cdots, M$$

Given the values of X_j and η_j, the model is a typical mixed model and thus the EM algorithm described before applies here. Detailed steps are given below.

Step 0: Set $t = 0$ and initialize all parameters $\theta = \theta^{(t)}$.

Step 1: Calculate the posterior mean and posterior variance for β_j, $\forall j = 1, \cdots, M$, using

$$\hat{\beta}_j = E(\beta_j|\cdots) = \mu_\beta + \Sigma_\beta X_j^T V_j^{-1}(y_j - X_j\mu_\beta)$$

and

$$\hat{W}_j = \text{var}(\beta_j|\cdots) = \Sigma_\beta - \Sigma_\beta X_j^T V_j^{-1} X_j \Sigma_\beta$$

where

$$V_j = \text{var}(y_j|\eta_j) = X_j\Sigma_\beta X_j^T + Z[\eta_j\sigma_1^2 + (1-\eta_j)\sigma_0^2]Z^T + I\sigma^2$$

Step 2: Calculate the posterior mean and posterior variance for γ_j, $\forall j = 1, \cdots, M$, using

$$\hat{\gamma}_j = E(\gamma_j | \cdots) = \Theta_j Z^T V_j^{-1}(y_j - X_j \mu_\beta)$$

and

$$\hat{S}_j = \text{var}(\gamma_j | \cdots) = \Theta_j - \Theta_j Z^T V_j^{-1} Z \Theta_j$$

where

$$\Theta_j = \text{var}(\gamma_j | \eta_j) = \eta_j \sigma_1^2 + (1 - \eta_j)\sigma_0^2$$

Step 3: Update μ_β using

$$\mu_\beta = \left[\sum_{j=1}^{M} X_j^T V_j^{-1} X_j\right]^{-1} \left[\sum_{j=1}^{M} X_j^T V_j^{-1} y_j\right]$$

Step 4: Update Σ_β using

$$\begin{aligned} \Sigma_\beta &= \frac{1}{M}\sum_{j=1}^{M} E\left[(\beta_j - \mu_\beta)(\beta_j - \mu_\beta)^T\right] \\ &= \frac{1}{M}\sum_{j=1}^{M} \left[(\hat{\beta}_j - \mu_\beta)(\hat{\beta}_j - \mu_\beta)^T + \hat{W}_j\right] \end{aligned}$$

Step 5: Update σ_1^2 using

$$\begin{aligned} \sigma_1^2 &= \frac{1}{\pi_1 M + d_0}\left[\sum_{j=1}^{M} \eta_j E(\gamma_j^2) + \omega_0\right] \\ &= \frac{1}{\pi_1 M + d_0}\left[\sum_{j=1}^{M} \eta_j (\hat{\gamma}_j^2 + \hat{S}_j) + \omega_0\right] \end{aligned}$$

Step 6: Update σ^2 using

$$\begin{aligned} \sigma^2 &= \frac{1}{MN} E\left[\sum_{j=1}^{M} (y_j - X_j\beta_j - Z\gamma_j)^T (y_j - X_j\beta_j - Z\gamma_j)\right] \\ &= \frac{1}{MN}\sum_{j=1}^{M} \left[(y_j - X_j\hat{\beta}_j - Z\hat{\gamma}_j)^T (y_j - X_j\hat{\beta}_j - Z\hat{\gamma}_j) + Z^T Z \hat{S}_j + \text{tr}(X_j^T X_j \hat{W}_j)\right] \end{aligned}$$

Step 7: Increment t by one and repeat from Step 1 to Step 6 until a certain criterion of convergence is reached. Note again that Steps 1-2 represent the E-steps and Steps 3-6 represent the M-steps.

3.2.4 Stochastic expectation and maximization (SEM) algorithm

3.2.4.1 Multiple eQTL

The EM algorithms described previously depend on known values of $\eta_{jk}, \forall j = 1, \cdots, M \& k = 1, \cdots, p$. They are missing values also but can be sampled from their posterior distributions. Once these missing values are replaced by values generated from the stochastic process, the above EM algorithm is called the stochastic EM algorithm (Celeux and Diebolt, 1985). The stochastic step is not parallel to the EM steps but an extra step inserted in the EM steps. This section describes the stochastic process. Note that the prior distribution for η_{jk} is $p(\eta_{jk}) = \text{Bernoulli}(\eta_{jk}|\pi_k)$, but the posterior distribution is $p(\eta_{jk}) = \text{Bernoulli}(\eta_{jk}|\rho_{jk})$ where ρ_{jk} is the posterior mean of η_{jk} and defined as

$$\rho_{jk} = \frac{\pi_k N(y_j|1\mu_\beta, V_{jk})}{\pi_k N(y_j|1\mu_\beta, V_{jk}) + (1-\pi_k)N(y_j|1\mu_\beta, V_{j0})} \tag{3.3}$$

where

$$V_{jk} = \sigma_\beta^2 11^T + \sum_{k' \neq k} Z_{k'}[\eta_{jk'}\sigma_{k'}^2 + (1-\eta_{jk'})\sigma_0^2]Z_{k'}^T + \sigma_k^2 Z_k Z_k^T + I\sigma^2$$

and

$$V_{j0} = \sigma_\beta^2 11^T + \sum_{k' \neq k} Z_{k'}[\eta_{jk'}\sigma_{k'}^2 + (1-\eta_{jk'})\sigma_0^2]Z_{k'}^T + \sigma_0^2 Z_k Z_k^T + I\sigma^2$$

The sampled η_{jk} are then used to infer π_k, $\forall k = 1, \cdots, p$ with the following equation

$$\pi_k = \frac{1}{M}\sum_{j=1}^{M} \eta_{jk}$$

Incorporating this stochastic step into the EM steps, we conclude the SEM algorithm.

3.2.4.2 Gene-trait association

For the gene-trait association study, both X_j and η_j are missing and both are sampled in the stochastic process. The sampling process for η_j is the same as the one

described early. The prior distribution for η_j is $p(\eta_j) = \text{Bernoulli}(\eta_j|\pi)$, but the posterior distribution is $p(\eta_j) = \text{Bernoulli}(\eta_j|\rho_j)$ where ρ_j is the posterior mean of and defined as

$$\rho_j = \frac{\pi N(y_j|X_j\mu_\beta, V_{j1})}{\pi N(y_j|X_j\mu_\beta, V_{j1}) + (1-\pi)N(y_j|X_j\mu_\beta, V_{j0})}$$

where

$$V_{j1} = X_j\Sigma_\beta X_j^T + \sigma_1^2 ZZ^T + I\sigma^2$$

and

$$V_{j0} = X_j\Sigma_\beta X_j^T + \sigma_0^2 ZZ^T + I\sigma^2$$

The sampled η_j is then used to infer π, as given below,

$$\pi = \frac{1}{M}\sum_{j=1}^{M}\eta_j$$

Since X_j is also missing, stochastic sampling is required to generate X_j. Recall that $X_j = [X_j^{(1)}, X_j^{(2)}]$ and $X_j^{(1)} + X_j^{(2)} = 1$, only the first column is sampled. We propose to sample $X_j^{(1)}$ one element at a time, conditional on values of all other elements. Let us now focus on the sampling of $X_{ji}^{(1)}$ for $i = 1, \cdots, N$ whose prior distribution is $p(X_{ji}^{(1)}) = \text{Bernoulli}(X_{ji}^{(1)}|\phi_j)$, where ϕ_j is the proportion of individuals who should be modeled with intercept β_{j1} for gene j for $j = 1, \cdots, M$. Let $H_{ji}^{(1)}$ be matrix X_j with the ith row replaced by $[1, 0]$. Similarly, define $H_{ji}^{(2)}$ as matrix X_j with the ith row replaced by $[0, 1]$. The posterior distribution for $X_{ji}^{(1)}$ is now $p(X_{ji}^{(1)}|\cdots) = \text{bernoulli}(X_{ji}^{(1)}|\rho_{ji})$. The posterior mean ρ_{ji} is inferred from the following Bayes' theorem,

$$\rho_{ji} = \frac{\phi_j N(y_j|H_{ji}^{(1)}\mu_\beta, V_j^{(1)})}{\phi_j N(y_j|H_{ji}^{(1)}\mu_\beta, V_j^{(1)}) + (1-\phi_j)N(y_j|H_{ji}^{(2)}\mu_\beta, V_j^{(2)})}$$

where

$$V_j^{(1)} = H_{ji}^{(1)}\Sigma_\beta H_{ji}^{(1)T} + Z[\eta_j\sigma_1^2 + (1-\eta_j)\sigma_0^2]Z^T + I\sigma^2$$

And

$$V_j^{(2)} = H_{ji}^{(2)}\Sigma_\beta H_{ji}^{(2)T} + Z[\eta_j\sigma_1^2 + (1-\eta_j)\sigma_0^2]Z^T + I\sigma^2$$

Once $X_j^{(1)}$ is sampled, we let $X_j^{(2)} = 1 - X_j^{(1)}$ to form a complete matrix X_j. The nuisance parameter ϕ_j is updated using

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} X_{ji}^{(1)}$$

3.2.4.3 Convergence criterion

With the SEM algorithm, parameters do not converge to some constant values; rather they converge to a joint stationary distribution, much like the MCMC algorithm. However, the SEM algorithm reaches to the stationary distribution much quicker than the MCMC algorithm because only a subset of the parameters are subject to sampling. The convergence can be visualized by the trace plot for each parameter. Once the stationary distribution is achieved for every parameter, the parameter values are collected for a period of times. The average values of the parameters over the iterations (after convergence) are the SEM estimates of the parameters. A more rigorous approach is to calculate the mean of each parameter for consecutive S iterations and monitor the convergence of the means. The mean of parameter vector θ at iteration t is the average value of θ over the last S iterations prior to iteration t, defined as

$$\overline{\theta}^{(t)} = \frac{1}{S} \sum_{s=1}^{S} \theta^{(t+1-s)}$$

Of course, the mean vector is only calculated after $t = S$. The convergence criterion is defined as

$$\|\overline{\theta}^{(t)} - \overline{\theta}^{(t-1)}\| \leq \delta$$

where $\delta = 10^{-4}$ or any other small number predefined by the investigator.

3.3 Application

3.3.1 Simulation study

3.3.1.1 Multiple eQTL

In this simulation experiment, ten markers ($p = 10$) were evenly placed on a 360-cM genome, with 40 cM distance per marker interval. Among the ten markers, four of them were assigned eQTL effects, which are marker 1 (0 cM) , marker 3 (80 cM) , marker 6 (200 cM) and marker 10 (360 cM). We simulated $N = 100$ individuals from a F2 family. A total of $M = 1000$ transcripts were simulated, among which transcripts 605-610 (six transcripts) were affected by eQTL at marker 1, transcripts 601-604 (four transcripts) were affected by eQTL at marker 3, transcripts 961-1000 (40 transcripts) were affected by eQTL at marker 6, and transcripts 1-50 (50 transcripts) were affected by eQTL at marker 10. The total number of transcripts controlled by the eQTL was $6 + 4 + 40 + 50 = 100$. Intercepts for the 1000 transcripts were randomly sampled from $U(\beta_j|2, 4)$. Each of the 100 eQTL effects was simulated from a $N(\gamma_{jk}|0, 3^2)$ distribution. The residual errors for each transcript were simulated from a multivariate $N(\varepsilon_j|0, 0.1^2 \times I_{100})$ distribution for all $j = 1, \cdots, M$. The simulation experiment was replicated 20 times. Each of the 20 simulated samples was analyzed using two methods: (1) the MCMC implemented Bayesian method developed by Jia and Xu (2007); (2) the SEM algorithm proposed here. In both analyses, the hyper parameters were set at $(d_0, \omega_0) = (5, 50)$ according to the parameters used in Jia and Xu (2007). The results are given in Table 3.1 and depicted in Figures 3.1 and 3.2. Both algorithms provided fairly accurate estimates of the proportions of transcripts associated with the markers and comparably estimates of σ_k^2 (see Table 3.1). However, the standard deviation of σ_k^2 estimated by SEM algorithm are much smaller than MCMC (see Table 3.1). Figure 3.1 shows the true and estimated

proportions of the transcripts controlled by the ten simulated markers, from which we can see that the estimated proportions are very close to the true proportions. We used $E(\eta_{jk}|\cdots) = \rho_{jk} > 0.9$ as the criterion of detection of the eQTL effect. Among the 100 eQTL effects, 99 of them were detected. The one failed to be detected was transcript 19 linked to marker ten with a true effect of 0.036 (too small to be detected). The true and estimated effects from both algorithms are shown in Figure 3.2. Overall, both SEM and MCMC gave satisfactory estimates of the parameters, especially the proportions of linked transcripts (very important parameters of the experiment). The computing times of these two methods were drastically different. The SEM algorithm only took 15 minutes to finish the analysis of each replicate whereas the MCMC algorithm took more than four hours to finish.

Table 3.1: Average mean and standard deviation(in parentheses) of π_k and σ_k^2 in 20 replicates of model I and Ebayes in simulation 1.

Marker Number		π_k	σ_k^2
1	True	0.006	9
	SEM	0.006 (1.5e-4)	7.49 (0.091)
	MCMC	0.007 (1.6e-4)	9.17 (0.267)
3	True	0.004	9
	SEM	0.004 (3.0e-5)	9.65 (0.033)
	MCMC	0.005 (1.5e-4)	12.32 (0.474)
6	True	0.040	9
	SEM	0.041 (3.0e-4)	11.30(0.070)
	MCMC	0.041 (4.6e-4)	11.88 (0.177)
10	True	0.050	9
	SEM	0.050 (3.7e-4)	9.25 (0.065)
	MCMC	0.050 (6.1e-4)	9.69 (0.148)

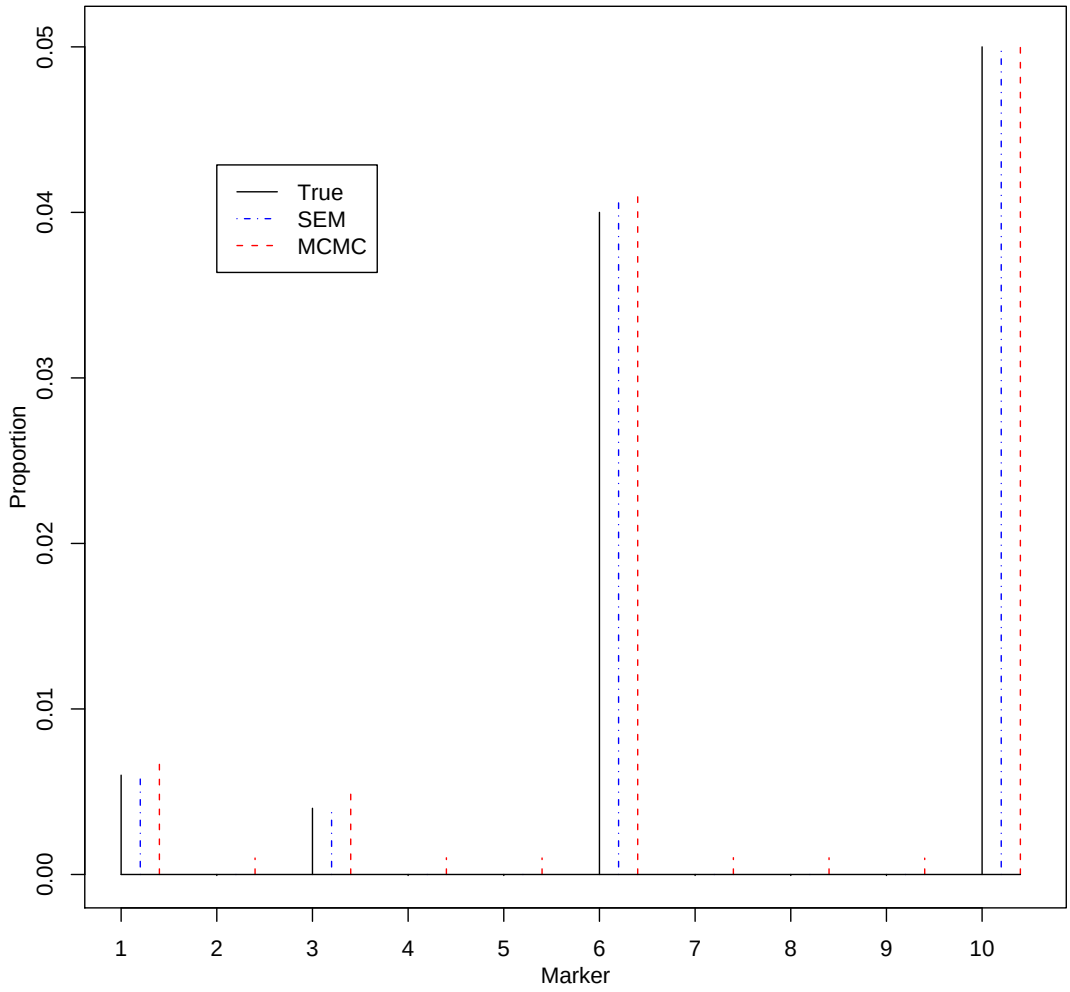

Figure 3.1: True and estimated proportions of associated transcripts for ten markers in the replicated multiple eQTL simulation experiment. The true proportions are indicated by the solid black vertical lines. The estimated proportions from the SEM are indicated by the dot-dashed blue vertical lines. The estimated proportions from the MCMC algorithm are indicated by the dashed red vertical lines.

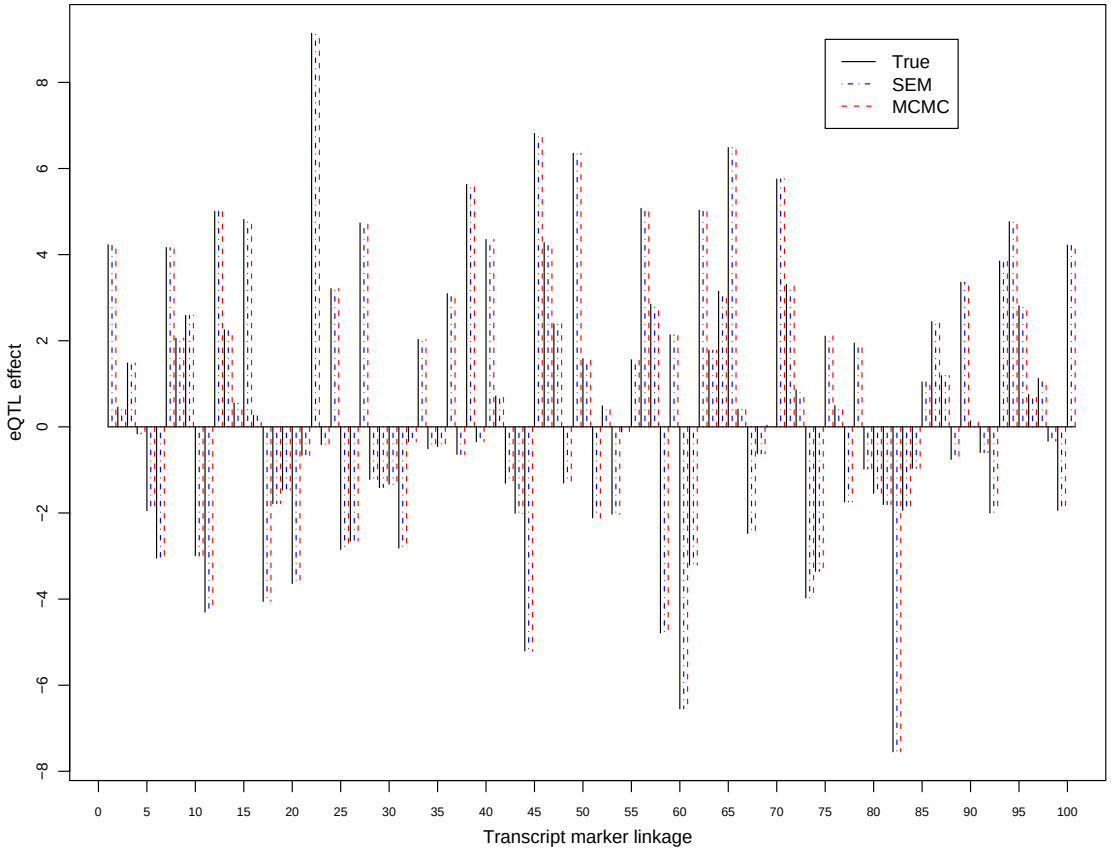

Figure 3.2: True and estimated eQTL effects for the 100 transcripts obtained from 20 replicated simulation experiment under the multiple eQTL model. The true effects are indicated by the solid black vertical lines. The estimated effects from the SEM are indicated by the dot-dashed blue vertical lines. The estimated effects from the MCMC algorithm are indicated by the dashed red vertical lines.

3.3.1.2 Gene-trait association with one intercept

In the second simulation experiment, the data were simulated under the single intercept model but the analyzes were conducted under two models, one of which was the single intercept model (called SEM I) and the other was the two-intercept model (called SEM II). We simulated 100 subjects and 1000 transcripts among which transcripts 1-50, 601-610, and 961-1000 were affected by the quantitative trait. The total number of associated transcripts was $50 + 10 + 40 = 100$. The phenotypic values of the quantitative trait were sampled from $U(-1, 1)$ for each of the 100 individuals. The intercept β_j was simulated from $U(\beta_j|2, 4)$ with an average value of $\mu_\beta = 3.0$. The regression coefficients γ_j were simulated from $N(\gamma_j|0, 3^2)$. The residual errors were simulated from a multivariate $N(\varepsilon_j|0, 0.1^2 I_{100})$ distribution. The experiment was replicated 20 times. Using the $\rho_j > 0.9$ criterion, both SEM I and SEM II detected at least 99 genes (out of 100) that are associated with the trait. The one that failed to be detected in some replicates was gene number 24 with a true effect of association of 0.059 (very small effect). There are no false detected genes in both algorithms. The estimated proportions of genes associated with the trait were 0.1012 and 0.1013, respectively for SEM I and SEM II. These estimated proportion are very close to the true value of $\pi = 0.10$. The true and estimated effects for the 100 genes are depicted in Figure 3.3. The conclusion was that when the data were simulated from the single intercept model, both the single intercept and two-intercept models were effective and the two-intercept model was robust to the assumption of the data structure.

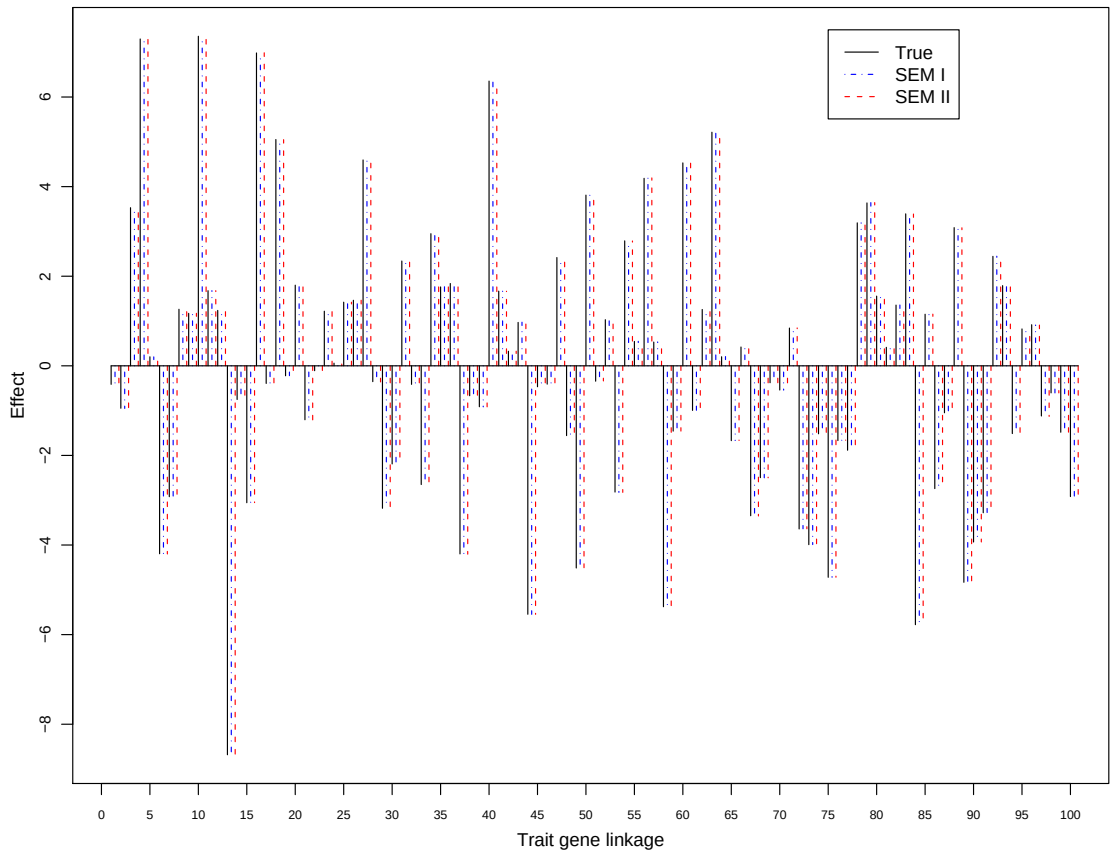

Figure 3.3: True and estimated effects for the 100 genes associated with the trait obtained from 20 replicated simulation experiment under the gene-trait association model. All genes were simulated from the single intercept model. The true effects are indicated by the solid black vertical lines. The estimated effects from SEM I (single intercept) are indicated by the dot-dashed blue vertical lines. The estimated effects from SEM II (two-intercept) are indicated by the dashed red vertical lines.

3.3.1.3 Gene-trait association with two intercept

In the third simulation experiment, we kept everything the same as in the second simulation except that some of the genes were simulated with two intercepts. Among the 1000 simulated genes, 100 genes were associated with the trait and the effects were simulated from $N(\gamma_j|0, 3^2)$. Among the 1000 genes, 200 genes were simulated with two intercepts and the remaining 800 genes were simulated with a single intercept for each gene. For the single intercept genes, the intercept was simulated from a $U(\beta_j|2, 4)$ distribution. For the 200 two-intercept genes, the first intercept was simulated from $U(\beta_{j1}|2, 4)$ and the second intercept was simulated from $U(\beta_{j2}|4, 6)$. Therefore, the expectation of the two intercepts were $\mu_1 = 3$ and $\mu_2 = 5$, respectively. For the 200 genes with two intercepts, the allocations of the subjects to the first and second intercepts varied. Some genes split the 100 subjects into 20 (β_{j1}) and 80 (β_{j2}), some split into 40 (β_{j1}) and 60 (β_{j2}), and others split into 60 (β_{j1}) and 40 (β_{j2}). Again, the simulation was replicated 20 times. The two models used to analyze the data were SEM I (single intercept model) and SEM II (two-intercept model). The criterion for detection of an associate remained $\rho_j > 0.9$.

In the SEM I analysis, the number of genes had $\rho_j > 0.9$ is between 171 and 178 according to different replicates. However, only 94-95 genes were truly associated and the remaining genes were false positive (See Table 3.2 for details). All the false positive genes belonged to the 200 two-intercepted genes. The estimated proportion of associated genes was $\hat{\pi} = 0.196$, much higher than the true proportion of 0.10. Recall that 200 genes were simulated with two intercepts, but the model did not have the ability to handle the two-intercept genes. This explains the lost power and gained false positive rate. Here the number of associated genes based on proportion $\hat{\pi}$ is a little bit larger than

the number of selected genes by criterion $\rho_j > 0.9$. The reason is that the criterion for detection of an associate is very strictly and eliminates some associations with $\rho_j \leq 0.9$. While estimated proportion $\hat{\pi}$ accounts for all non-zero ρ_j.

In the SEM II analysis, a total number of genes which were claimed to be associated with the trait is between 99 to 102 according to different replicates. Among the detected genes, at least 99 were truly associated and at most one was false positive (See Table 3.2 for details). The estimated proportion of associated genes was $\hat{\pi} = 0.103$, very close to the true value of 0.10. Recall again that 200 genes were simulated with two intercepts. The model allowed two intercepts to occur for some genes, and thus was able to reduce the false positive rate. The true and estimated effects for the 100 genes are depicted in Figure 3.4.

Table 3.2: Number of detected genes by SEM I and II in simulation 3. Column two and four represent the total number of detected genes by SEM I and II. Column three and five represent the number of true associated genes among detected genes.

	SEM I		SEM II	
Replicate	Detected genes	True genes	Detected genes	True genes
1	178	95	100	99
2	174	95	100	99
3	173	95	101	100
4	177	95	102	100
5	175	95	101	99
6	173	94	100	99
7	175	95	101	99
8	174	95	100	99
9	174	95	100	99
10	178	95	101	99
11	174	95	100	99
12	172	95	101	99
13	172	94	102	99
14	173	94	99	99
15	173	95	100	99
16	173	95	99	100
17	174	95	100	99
18	174	95	100	99
19	174	94	100	99
20	171	95	99	99

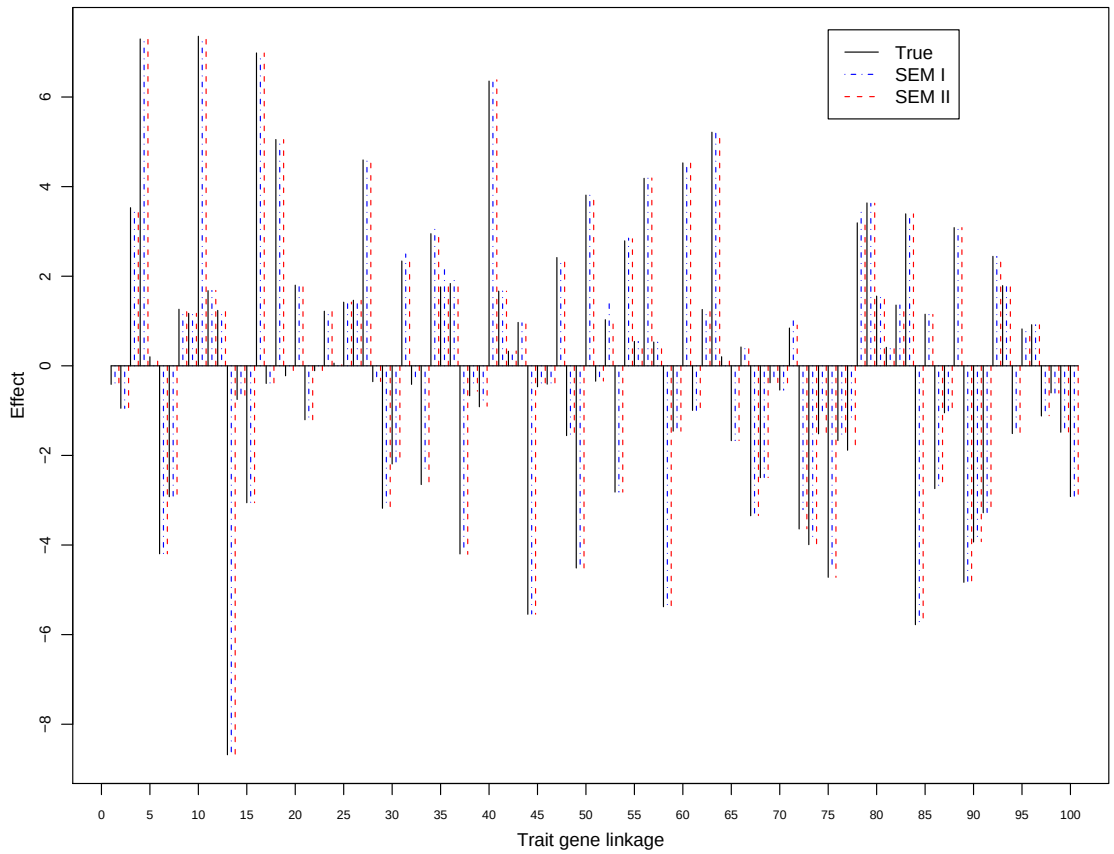

Figure 3.4: True and estimated effects for the 100 genes associated with the trait obtained from 20 replicated simulation experiment under the gene-trait association model. Among the 1000 genes, 200 of them were simulated from the two-intercept model. The true effects are indicated by the solid black vertical lines. The estimated effects from SEM I are indicated by the dot-dashed blue vertical lines. The estimated effects from SEM II are indicated by the dashed red vertical lines.

3.3.2 Real data analysis

We apply our method to barley data to find the linkage between gene expression, quantitative traits, and biomarker data. The procedure is carried on as follows: we first use model II to pick up transcripts with strong association with quantitative traits and then apply model I to do linkage study between selected transcripts and markers.

The gene expression data were published by Luo et al. (2007) and downloadable from the following website: http://www.ebi.ac.uk/microarray-as/aer/entry. The phenotypic values of eight quantitative traits of barley were published by Hayes et al. (1993) and downloadable from the following website: http://wheat.pw.usda.gov/ggpages/SxM/phenotypes.html. Detailed description of the experiment can be found from the original study (Hayes et al., 1993). The experiment involved 150 double haploid (DH) lines derived from the cross of two spring barley varieties, Morex and Steptoe. All the 150 DH lines were microarrayed for 22840 transcripts. The eight traits are alpha amylase, disastatic power, grain protein, heading date, height, lodging, malt extract, and yield. The phenotypes of the traits were measured in different environments (locations and years). The number of replicated measurements ranged from 6 to 16 depending on different traits. We took the average of replicates to stand for different traits. We analyzed eight traits by model I(single marker case) and II. In model II, we need to assign two intercepts to each transcript to account for clustering. However, it might be hard to give prior μ_1, μ_2 due to different scale of Microarray data. For example, there are two clustered transcripts. Two cluster means for the first transcript are 2 and 4 and the other are 4 and 6. In this case it is hard to choose prior μ_1, μ_2 to cluster both two transcripts. If we use prior $\mu_1 = 2, \mu_2 = 4$, we can cluster first transcript. But for the second transcript it is impossible to do the clustering since $\mu_2 = 4$ is more close to whole

data. There might be no transcripts in first cluster and all transcripts in second cluster. So before we applied model II, we rescaled Microarray data. The subjects of each transcripts were subtracted by the minimum value of that transcripts, which makes sure the first cluster of each transcript with cluster mean close to 0. In order to compare the results of model I and II, we first use k-mean method to cluster each transcripts into two groups. Then we selected transcripts with huge difference between two cluster means as references to compare model I and II. Totally we selected 279 transcripts with cluster mean differences greater than 2 and computed how many of these clustered transcripts were detected by model I and II with linkage estimator $\overline{\eta}_j$ greater than 0.9. Since model I did not consider two intercepts, it may have more chance to detect these clustered transcripts than model II. We can clearly see this phenomenon from figure 3.5 and 3.6 which represents eight clustered transcripts detected by model I instead of model II for eight traits. The first column represents original data and the second column represents adjusted data by model II. From the original data, the detected transcripts have clearly two clusters which are almost parallel with each other. They shows strong linear correlation due to two parallel clusters. However, the transcripts have almost no correlation with traits if we adjust the intercepts for two clusters. Figure 3.5 and 3.6 only showed one transcript for each traits. Actually there are a lot of clustered transcripts detected by model I. Table 3.3 shows the total number of detected transcripts in model I and II, the number of detected transcripts belongs to 279 clustered transcripts by model I and II. From the table the number of clustered transcripts detected by model I is almost twice than that in model II.

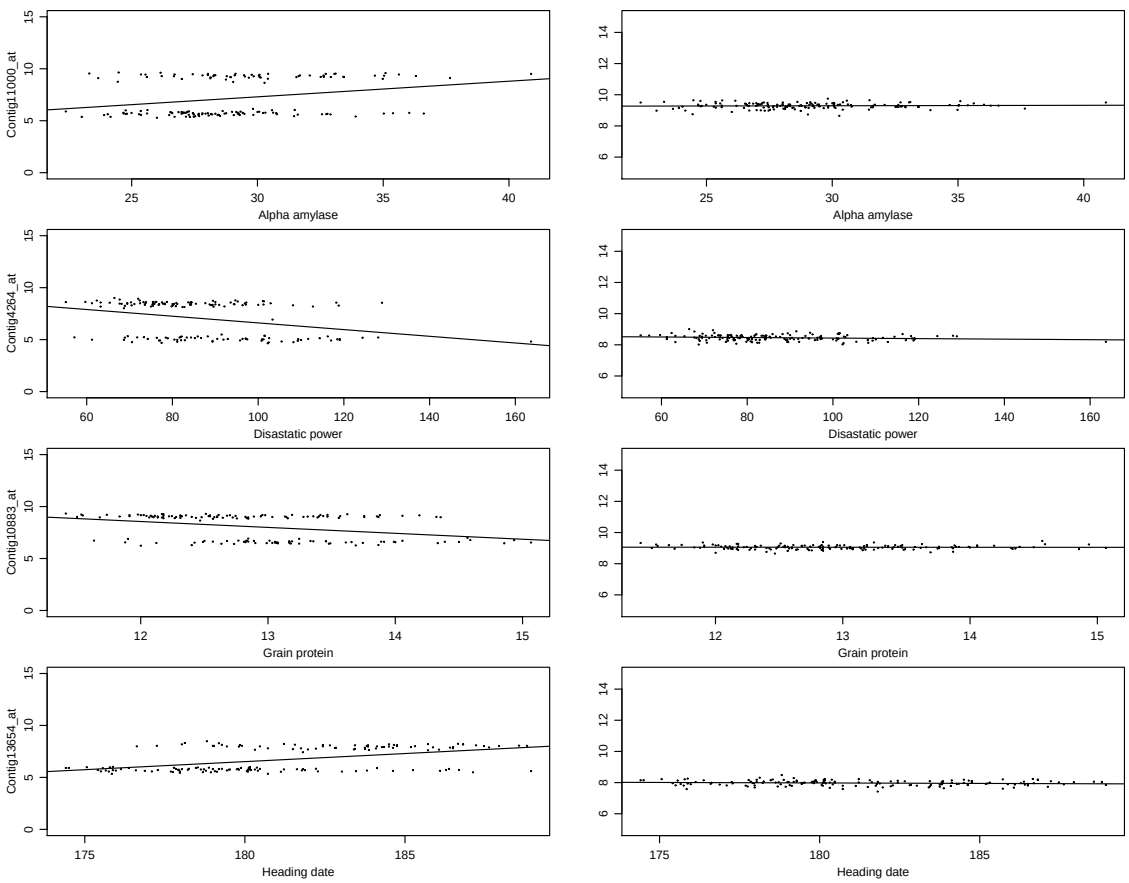

Figure 3.5: Four selected transcripts detected by model I for first four traits. First column represents linear relationship between transcripts and traits based on original data. Column two represents linear correlation between adjusted transcripts and traits by model II.

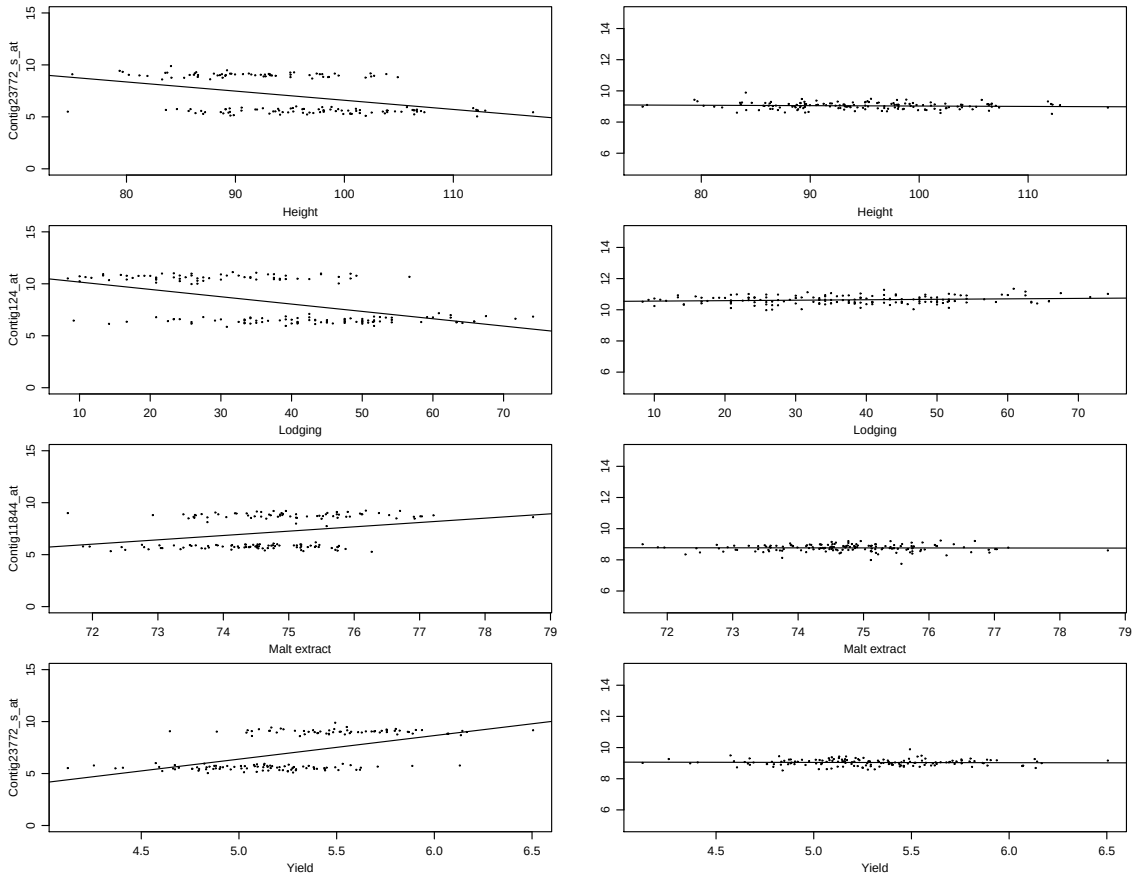

Figure 3.6: Four selected transcripts detected by model I for last four traits. See figure 3.5 for legends.

Table 3.3: Number of transcripts associated with 8 traits in model I and II. The first column in model I and II represent total number of detected transcripts. The second column represent number of detected transcripts belonging to 279 clustered transcripts.

Quantitative	Model I		Model II	
traits	Detected	Clustered	Detected	Clustered
Alpha amylase	605	156	973	35
Disastatic power	440	148	509	37
Grain protein	866	185	1434	55
Heading date	385	103	564	28
Height	768	154	1177	45
Lodging	626	175	859	40
Malt extract	503	175	576	27
Yield	434	130	669	36

Our next step is to do eQTL mapping. There are total 495 markers with an average marker interval less than 2 centiMorgan in the barley data. We applied model I to consider 495 markers simultaneously. It is hard to use Ebayes method to analyze barley data due to a huge number of markers. So in eQTL analysis, we only use model I to analyze the data. Figure 3.7 showed the proportion of transcripts associated with markers for eight traits. From the figures, there is barely any association between transcripts and markers in chromosome 1, 4, 6, 7 for all eight traits. The hottest spot is in chromosome 2.

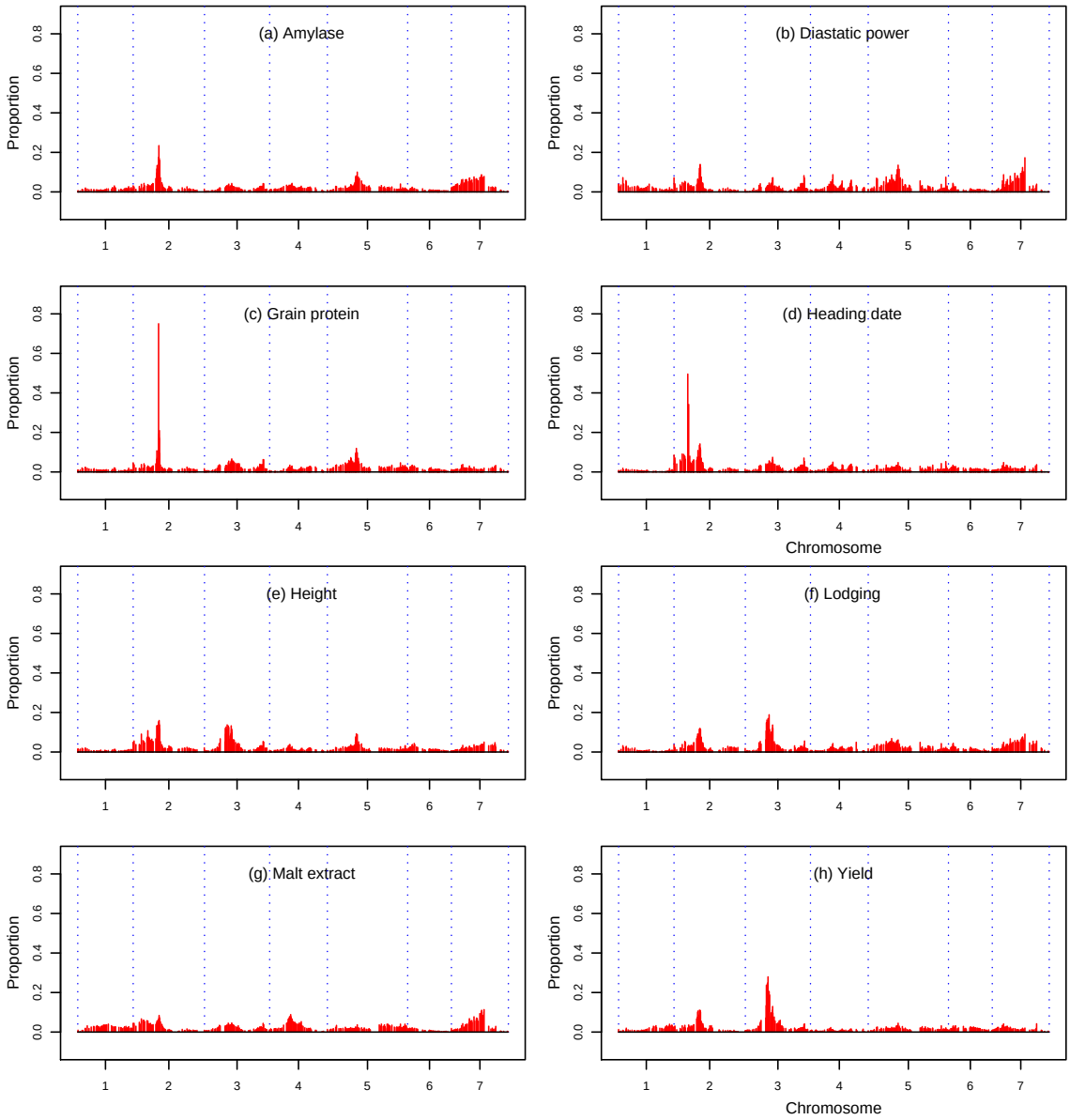

Figure 3.7: Marker transcript linkage map of eight traits for barley data. The transcripts were based on model II. Red vertical lines represent the proportion of transcripts controlled by a marker. Chromosome are separated by vertical dotted reference lines.

3.4 Discussion

We proposed SEM algorithm to analyze eQTL data and Quantitative trait associated Microarray data. The SEM algorithm is a hybrid of EM algorithm and MCMC algorithm. The computing time of SEM algorithm is dramatically decreased compared with MCMC algorithm since not all variables are sampled to generate posterior distribution in MCMC. Meanwhile it will solve EM local maximization problem. In our study, the results of simulation 1 demonstrated the performances of SEM and MCMC were almost same. However, the computing time of MCMC was much longer than SEM. In fact the computing time problem will be more obvious and severe in real data analysis. With the development of biotechnology, the gene expression data and SNP data will grow larger and larger. It is impossible for MCMC algorithm to analyze tens of thousands of data even with super powerful computer. We think SEM algorithm is a good alternative. Actually more and more researchers realized time consuming problem and try to use more efficient algorithms. The original Bayesian shrinkage analysis of Xu (2003) implemented through MCMC algorithm was improved by Xu (2010) through EM algorithm. The improved method achieved the same goal as MCMC algorithm but with much shorter time and was comparable with Lasso (Tibshirani, 1996).

The gene-trait association study is a developing area, which focuses on finding correlation between gene expression and a continuous phenotype. The previous method proposed by Zhan et al. (2010) is able to find some transcripts at the cost of mixing with many differentially expressed transcripts. From figure 3.5 and 3.6, we cannot see strong linear correlation for these clustered transcripts. In fact, the clustered transcripts can exist in almost all Microarray data. The clustered transcripts can cause severe detection problem if we do gene-trait association study. Our proposed SEM model is the first

method aim to solve this problem.

The SEM algorithm can also be applied to differentially expression analysis. In this analysis, there are usually two conditions: treatment and control. So we can change matrix Z to be 0 or 1 corresponding to control or treatment. It is also a single marker model. The Gaussian mixture model can also play a similar role as that in eQTL and trait-gene association study.

In the chapter, we also provide a framework on the linkage study of three different datasets: gene expression data, quantitative trait, and snp data. The gene-trait association study is like a gene selection procedure to select most interested genes. Then based on selected genes, we do eQTL study to determine the linkage study between genes and snps. The advantage of this framework is time consuming. It might cost us plenty of time to do eQTL analysis of whole gene expression data and snp data. However, we may only interest in a few transcripts related to some particular research such as an important disease in humans or an economically important trait in agricultural species. In such case, it is not necessary to do eQTL study of whole gene expression data. Our proposed framework is one of the efficient ways to save time.

Chapter 4

eQTL analysis in deep sequencing

4.1 Introduction

In recently years, the development of high-throughout DNA sequencing technologies generates second-generation deep sequencing arrays. These technologies enable thousands of megabases of DNA to be sequenced in a matter of days. Currently there are three main technologies developed by 454 Life Sciences (Roche) (Margulies et al., 2005), Illumina (formerly Solexa sequencing) (Bennett et al., 2005), and ABI (SOLiD sequencing) (Shendure et al., 2005). For each of these sequencing platforms, a huge and complex data set is generated and requires many pre- and post- statistical methods to do the analysis. The first issue is base-call procedure. Several papers deal with this problem such as Erlich et al. (2008), Rougemont et al. (2008), and Irizarry and Bravo (2009). These methods mainly focus on millions of short nucleotide sequences, referred to as reads, which are strings of A,C,G or Ts between 30-100 characters long and try to improve the accuracy of converting raw intensities into discrete base calls. Once the sequence of data is determined, the next coming issue is the mapping of the (short) reads to the genome from which they derive. This procedure is also considered as align-

ment. Faulkner et al. (2008) employed a multi-mapping tag rescue strategy leading to a significant increase in mapping accuracy compared to previous methods (Lassmann et al., 2008). After alignment, we get mapped DNA sequence reads of the entire genome. When we have multiple samples we will have a read-count profile that counts the number of reads from each sample. However, there are some systematic variations between samples which might cause the read-count profile between samples incomparable. The normalization procedure (Balwierz et al., 2009) is a method to deal with this issues.

Although deep sequencing technologies have originally been used for genomic sequencing, more recently researchers have applied technologies for a number of other applications. we can get our gene expression profile based on above preprocessing analysis pipeline. Then high-level analysis can be applied to the profile to get biological or clinical applications such as gene differential expression analysis (Robinson and Smyth, 2007; Marioni et al., 2008). However, we still haven't found any paper to deal with the linkage problem between deep sequencing gene expression data with other types of data such as SNP data. In chapter 3, we proposed SEM algorithm to do eQTL study between Microarray gene expression data and marker data. Since there are some similarities between Microarray gene expression data and deep sequencing gene expression data, the linkage study here can be considered as an extension of eQTL study. We name this linkage study as deep sequencing eQTL study.

In traditional eQTL study, the gene expression data can be treated as normal distribution. So our proposed model in chapter 3 is based on linear model. The structure of deep sequencing gene expression data is quite different from Microarray gene expression data. The expression value for each gene is the number of counts, which is discrete data instead of continuous data. Usually, researchers consider deep sequencing gene expression data as poisson distribution. So We need to build a generalized linear model

for these data. However, it is difficult for classical generalized linear models to handle our proposed model in chapter 3. Since we have a lot of random effects in our model, it is impossible to integrate them out. We might use numerical method such as Laplace approximation to achieve integration. However, this process might be time consuming. Wolfinger and Oconnell (1993) proposed pseudo-likelihood approach is a fast and easy way to solve our problem. German et al. (2003); Gelman et al. (2008) extended pseudo-likelihood approach and give us a framework. The basic idea is to approximate the generalized linear model by a normal linear model and then apply the algorithm for normal linear models to estimate the parameters. In this chapter, we will first apply the above method to our proposed model and combine SEM algorithm presented before to estimate the parameters. We did two simulation studies to test the performance of our method. The results of simulation are very promising. We cannot apply our method to real data since it is impossible to find such data by far. However, we believe, our proposed method is a potential method to solve relevant linkage problems.

4.2 Method

In generalized linear model framework, we need to change our proposed model in chapter 3. We use the same index notation as in chapter 3. Let y_{ij} for $j = 1, \ldots, M$ and $i = 1, \ldots, N$ be the expression value of the jth gene measured from ith subject of the mapping population. Let Z_{ik} be a genotype indicator variable for subject i and marker k for $k = 1, \ldots, p$. For each gene expression data, we have following poisson distribution and link function,

$$
\begin{aligned}
f(y_{ij}|\beta_j,\gamma_{ik}) &= \frac{\mu_{ij}^{y_{ij}}\exp(-\mu_{ij})}{y_{ij}!} \\
\lambda_{ij} &= \log\mu_{ij}, \\
\lambda_{ij} &= \beta_j + \sum_{k=1}^{p} Z_{ik}\gamma_{jk}
\end{aligned} \tag{4.1}
$$

Based on above poisson model and link function, we first generate pseudodata and pseudovariance by current estimated parameters $\hat{\beta}$ and $\hat{\gamma}$. The pseudodata y_{ij}^* and pseudovariance σ_{ij}^2 can be generated by,

$$
\begin{aligned}
y_{ij}^* &= \hat{\lambda}_{ij} - \frac{l^{'}(y_{ij}|\hat{\lambda}_{ij})}{l^{''}(y_{ij}|\hat{\lambda}_{ij})} \\
\sigma_{ij}^2 &= -\frac{1}{l^{''}(y_{ij}|\hat{\lambda}_{ij})}
\end{aligned} \tag{4.2}
$$

where $\hat{\lambda}_{ij} = \hat{\beta}_j + \sum_{k=1}^{p} Z_{ik}\hat{\gamma}_{jk}$, $l(y_{ij}|\lambda_{ij}) = \log f(y_{ij}|\beta_j,\gamma_{ik})$, $l^{'}(y_{ij}|\lambda_{ij}) = dl(y_{ij}|\lambda_{ij})/d\lambda_{ij}$, $l^{''}(y_{ij}|\lambda_{ij}) = d^2l(y_{ij}|\lambda_{ij})/d\lambda_{ij}^2$, and $\hat{\beta}$ and $\hat{\gamma}$ are the current estimate of β and γ, respectively.

So in our model,

$$
\begin{aligned}
l(y_{ij}|\lambda_{ij}) &= \lambda_{ij}y_{ij} - \exp(\lambda_{ij}) - \log(y_{ij}!) \\
l^{'}(y_{ij}|\lambda_{ij}) &= y_{ij} - \exp(\lambda_{ij}) \\
l^{''}(y_{ij}|\lambda_{ij}) &= -\exp(\lambda_{ij}) \\
y_{ij}^* &= \hat{\lambda}_{ij} - 1 + \frac{y_{ij}}{\exp(\hat{\lambda}_{ij})} \\
\sigma_{ij}^2 &= \exp(-\hat{\lambda}_{ij})
\end{aligned}
$$

The generated pseudodata y_{ij}^* will follow an approximate normal distribution with mean λ_{ij} and variance σ_{ij}^2. So we can write linear model upon the pseudodata y_{ij}^* and the definition of λ_{ij} as follows,

$$
y_{ij}^* = \beta_j + \sum_{k=1}^{p} Z_{ik}\gamma_{jk} + \varepsilon_{ij} \tag{4.3}
$$

where

$$\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$$

Then we can apply SEM algorithm to current pseudodata to estimate parameters β and γ. The total algorithm can be achieved by three steps:

1. Generate pseudodata y_{ij}^* and pseudovariance σ_{ij}^2based on current estimated $\hat{\beta}$ and $\hat{\gamma}$.

2. Use SEM algorithm to estimate β and γ based on current generated pseudodata y_{ij}^*.

3. Repeat steps 1-2 until satisfied iterations reach.

4.3 Simulation

We carried out similar simulation experiment as that in chapter 3 to analyze the performance of above method. In first experiment, gene marker linkage mapping information is the same as that of first experiment in chapter 3. The intercept and eQTL effects are slightly different. The intercept β_j for 1000 transcripts were randomly sampled from a sequence starting from 2 and ending at 5 with increment by 0.1. The eQTL effects for the 100 linked transcripts (γ_{jk}) were simulated from Normal($\gamma_{jk}; 0, 2^2$). We can calculate λ_{ij} by intercept and eQTL effects. Then we can get mean μ_{ij} by link function. Finally we generate data y_{ij} by their corresponding mean and poisson distribution. The experiment was replicated 20 times. Table 4.1 shows the total number of detected true linkages and the number of detected false linkages (in bracket) for each replicates. Overall speaking, the detecting power is very high for each replicate. The slight different detecting power for each replicate are due to different generated data. In chapter 3, error cannot impact a lot in each replicate, which leads to same detecting

rate for total 20 replicates. Here we generate data according to poisson distribution. The same mean μ_{ij} can generate quite different data, which means error can impact the detecting power for different replicates. But the detecting power are still very consistent for 20 replicates. We found the missed linkages are due to small eQTL effects which are less than 0.1. Meanwhile, the false detecting linkage only happens in replicate 7 and 13. There are only 1 or 2 false detecting linkages in these two replicates, which means the type I error is pretty small. The estimated average proportion (π_k) of transcripts associated with each of 10 markers is displayed in figure 4.1. The estimated effects of 100 true linkages are presented in figure 4.2.

In second simulation, the linkage map information is a little bit different from first simulation. we let the eQTL at marker 1 control transcripts 1C20 and transcripts 971C990 and let the eQTL at marker 3 control transcripts 17C20. The transcripts controlled by the eQTL at markers 6 and 10 remained the same as in the first experiment. The purpose of the second simulation experiment was to allow some transcripts to be controlled by more than one marker. Table 4.2 shows the total number of detected true linkages and the number of detected false linkages (in bracket) for each replicates. Figure 4.3 shows estimated average proportion (π_k) of transcripts associated with each of 10 markers. The estimated effects of 100 true linkages are presented in Figure 4.4. From table 4.2 and figure 4.3, the detecting power is very high with tolerable small type I error. The accuracy of estimated eQTL effects is also very high according to figure 4.4.

Table 4.1: The number of detected true and false linkages according to different thresholds in simulation 1. The number in bracket represents false detected linkages

Thresholds	0.6	0.7	0.8	0.9
Rep1	96	96	96	96
Rep2	95	95	95	94
Rep3	97	97	97	97
Rep4	97	96	96	96
Rep5	95	95	94	94
Rep6	97	97	97	96
Rep7	96(1)	96(1)	96(1)	95
Rep8	96	94	94	94
Rep9	95	95	95	95
Rep10	94	94	94	94
Rep11	97	97	97	95
Rep12	96	96	96	96
Rep13	96(2)	96(2)	94(1)	94(1)
Rep14	95	95	95	95
Rep15	96	96	95	95
Rep16	96	95	95	94
Rep17	96	96	96	96
Rep18	97	97	96	95
Rep19	95	95	95	95
Rep20	96	96	96	96

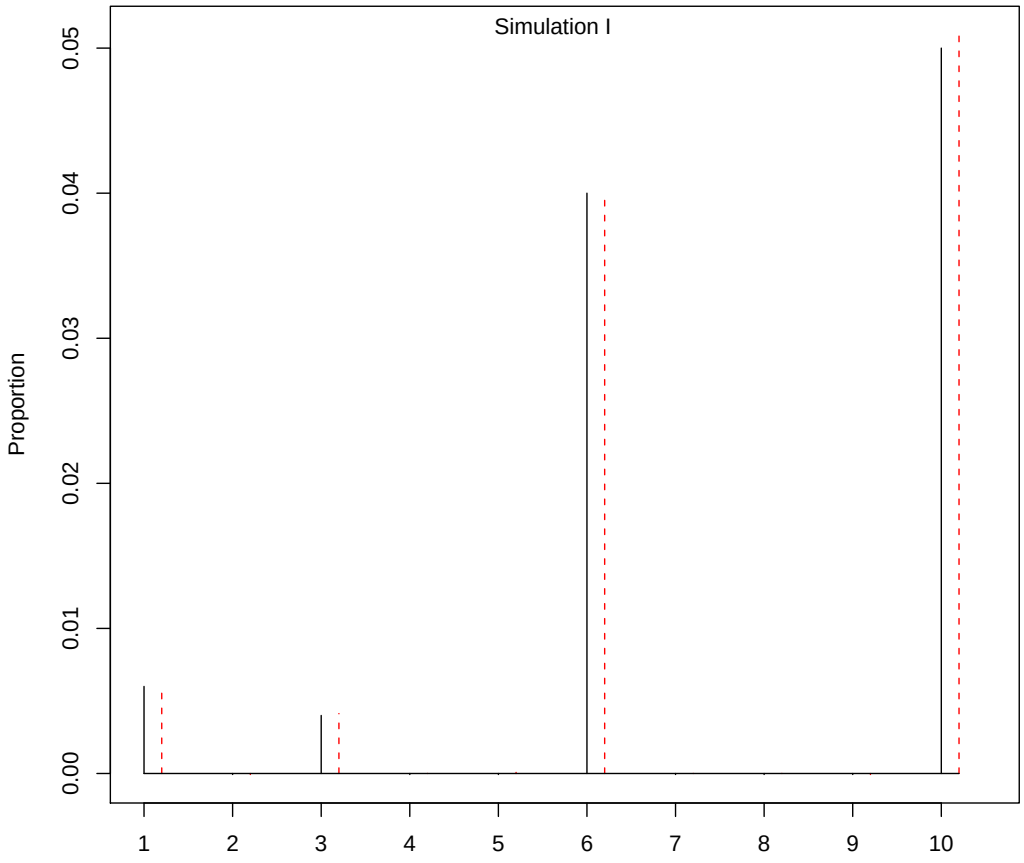

Figure 4.1: Proportion of associated transcripts for ten markers in simulation 1. Black vertical lines represent true proportion. Red dashed lines represent estimated proportion.

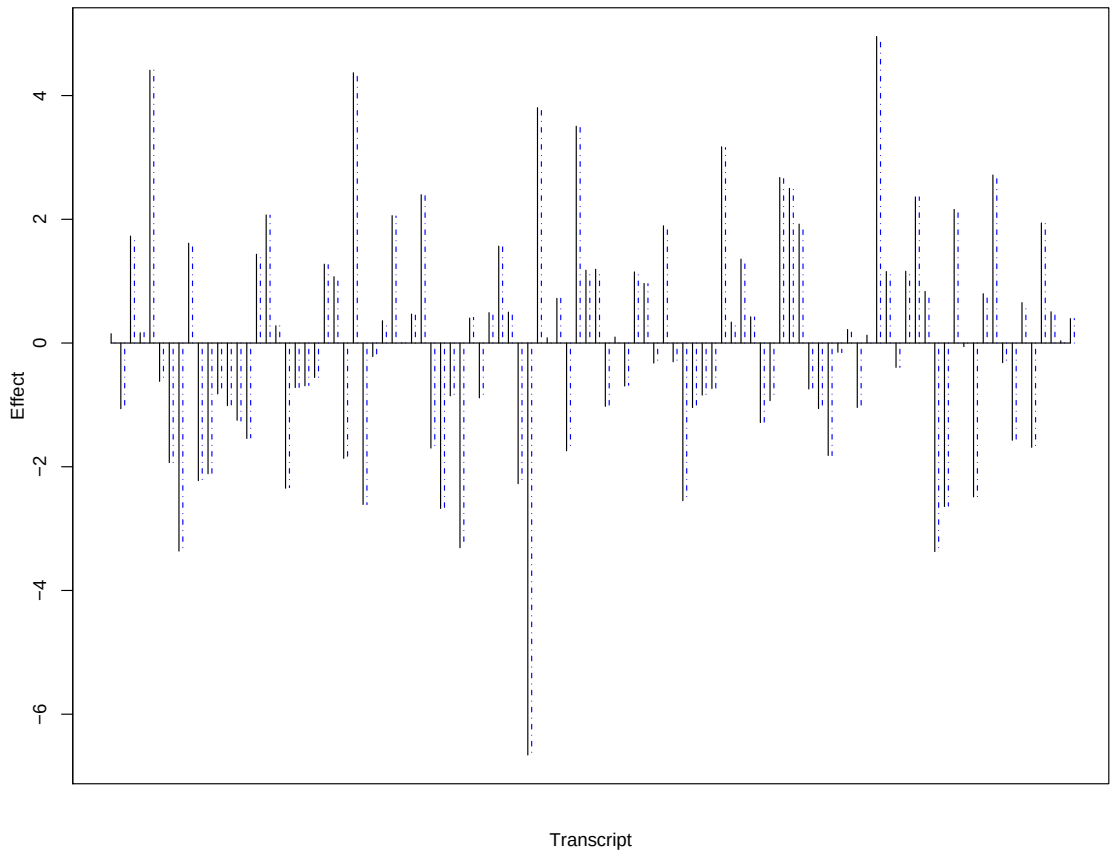

Figure 4.2: 100 true and estimated effects in simulation 1. Black vertical lines represent true effects. Blue dashed lines represent estimated effects.

Table 4.2: The number of detected true and false linkages according to different thresholds in simulation 2. The number in bracket represents false detected linkages

Thresholds	0.6	0.7	0.8	0.9
Rep1	131(2)	131(2)	131(2)	131(1)
Rep2	130(1)	130	130	130
Rep3	130(1)	130(1)	130(1)	130(1)
Rep4	131	131	131	131
Rep5	131	131	131	131
Rep6	131	131	131	131
Rep7	130	130	130	130
Rep8	131	130	130	130
Rep9	130	130	130	130
Rep10	130	130	130	130
Rep11	132	132	130	130
Rep12	131	131	131	131
Rep13	131	131	131	131
Rep14	130	130	130	130
Rep15	130	130	130	130
Rep16	131(1)	131	131	131
Rep17	131	131	131	130
Rep18	130	130	130	130
Rep19	131	131	131	131
Rep20	132	132	132	131

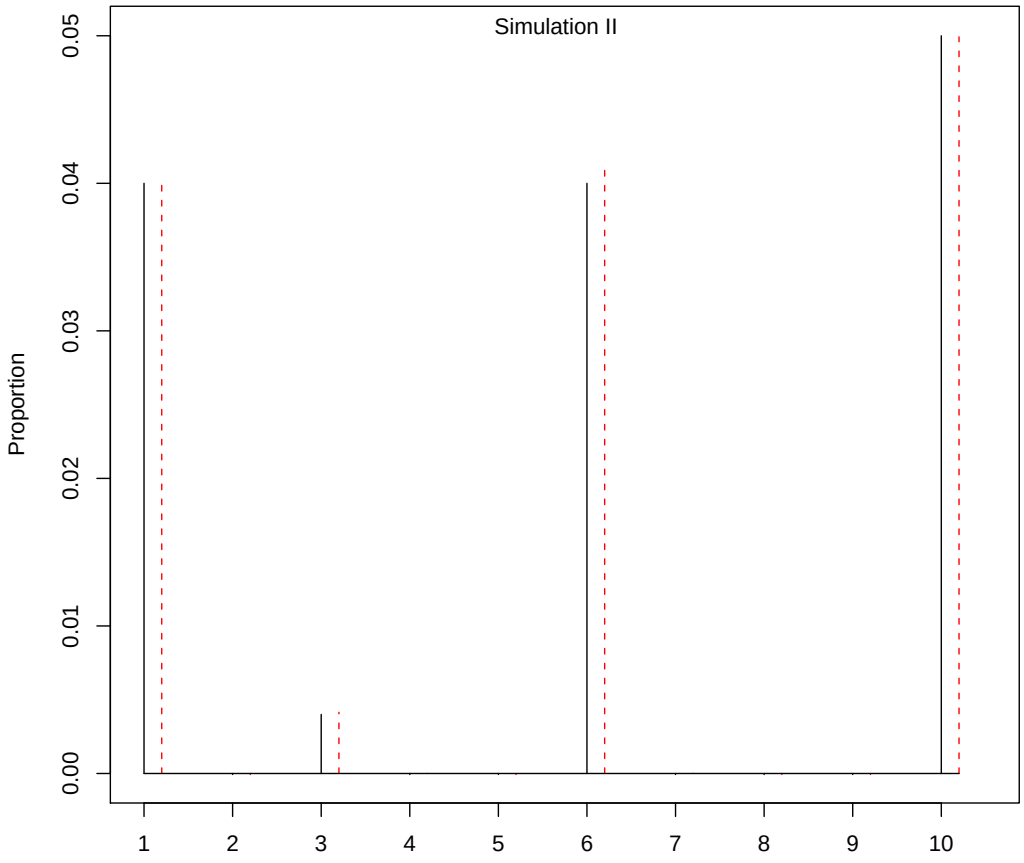

Figure 4.3: Proportion of associated transcripts for ten markers in simulation 2. Black vertical lines represent true proportion. Red dashed lines represent estimated proportion.

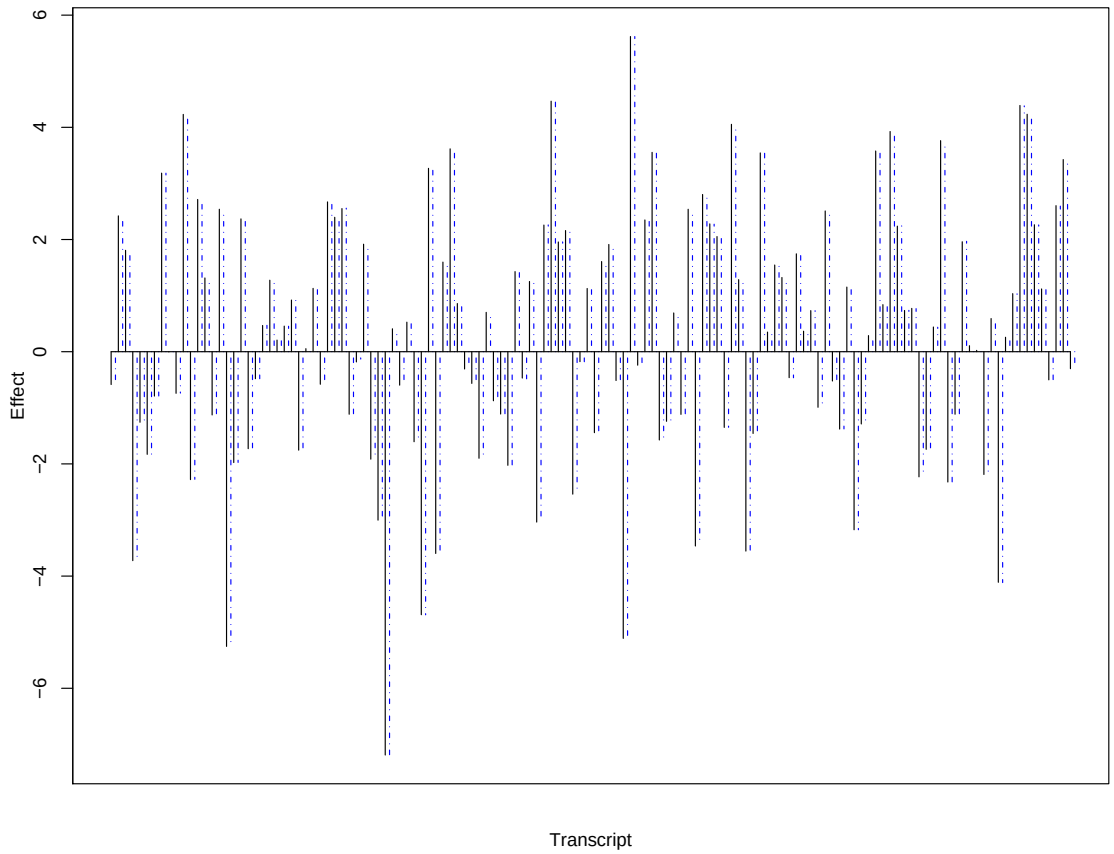

Figure 4.4: 134 true and estimated effects in simulation 2. Black vertical lines represent true effects. Blue dashed lines represent estimated effects.

4.4 Discussion

Our proposed method is the first method to deal with linkage problem between deep sequencing expression data and snp data. Due to different distribution assumption between Microarray gene expression data and deep sequencing expression data, it is impossible to implant eQTL method directly into deep sequencing eQTL problem. Pseudodata is like a bridge to overcome this obstacle. Based on SEM algorithm presented in Chapter 3, we can only include pseudodata generation procedure to the algorithm. The simulation studies demonstrate feasibility of proposed method. MOM is a very popular eQTL method which is also built on normal assumption. If expression data follow poisson distribution, they need to use conjugate prior gamma distribution to analyze the data, which might be time consuming than normal data. But still the assumption one gene only associates with at most one marker is a huge limitation.

Bibliography

D. Abdueva, D. Skvortsov, and S. Tavaré. Non-linear analysis of GeneChip arrays. *Nucleic Acids Research*, 34(15):e105, 2006.

Affymetrix. Affymetrix Microarray Suite Users Guide. Santa Clara, CA, version5.0 edition, 2001.

P. Baldi and A.D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17 (6):509, 2001.

P.J. Balwierz, P. Carninci, C.O. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. Van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol*, 10(7):R79, 2009.

S.T. Bennett, C. Barnes, A. Cox, L. Davies, and C. Brown. Toward the $1000 human genome. *Pharmacogenomics*, $6(4) : 373 - -382, 2005$.

B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262, 2000.

C.J. Burden, Y.E. Pittelkow, and S.R. Wilson. Statistical analysis of adsorption models for oligonucleotide microarrays. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 35, 2004.

C.J. Burden, Y. Pittelkow, and S.R. Wilson. Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays. *Journal of Physics: Condensed Matter*, 18:5545–5565, 2006.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

E. Chudin, R. Walker, A. Kosaka, S.X. Wu, D. Rabert, T.K. Chang, and D.E. Kreder. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol*, 3(1):0005.1–0005.10, 2001.

L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–331, 2004.

X. Cui and G.A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

J.L. Devore and R. Peck. Statistics: the exploration and analysis of data. Brooks, 1997.

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display

of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

Y. Erlich, P.P. Mitra, et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature methods*, 5(8):679–682, 2008.

G.J. Faulkner, A.R.R. Forrest, A.M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D.A. Hume, and S.M. Grimmond. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91(3):281–288, 2008.

A. Gelman, A. Jakulin, and M. Grazia. A weakly informative default prior distribution for logistic and other regression models. *Annals*, 2(4):1360–1383, 2008.

A. German, JB Carlin, HS Stern, and DB Rubin. Bayesian data analysis, 2003.

A. Halperin, A. Buhot, and EB Zhulina. Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophysical journal*, 86(2):718–730, 2004.

PM Hayes, BH Liu, SJ Knapp, F. Chen, B. Jones, T. Blake, J. Franckowiak, D. Rasmusson, M. Sorrells, SE Ullrich, et al. Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *TAG Theoretical and Applied Genetics*, 87(3):392–401, 1993.

X.J. He, Y.F. Hsu, S. Zhu, H.L. Liu, O. Pontes, J. Zhu, X. Cui, C.S. Wang, and J.K. Zhu. A conserved transcriptional regulator is required for RNA-directed DNA methylation and plant development. *Genes & development*, 23(23):2717, 2009.

D. Hekstra, A.R. Taussig, M. Magnasco, and F. Naef. Absolute mRNA concentrations

from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*, 31(7):1962–1968, 2003.

GA Held, G. Grinstein, and Y. Tu. Some M-estimates for expression analysis. *Affymetrix GeneChip microarray low-level workshop*, page http://www.affymetrix.com, 2003a.

GA Held, G. Grinstein, and Y. Tu. Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7575–7580, 2003b.

GA Held, G. Grinstein, and Y. Tu. Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Research*, 34(9):e70, 2006.

J. Herrero, A. Valencia, and J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126, 2001.

R.A. Irizarry and H.C. Bravo. Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, page 184, 2009.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15, 2003a.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003b.

R.A. Irizarry, Z. Wu, and H.A. Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7):789–794, 2006.

Z. Jia and S. Xu. Mapping quantitative trait loci for expression abundance. *Genetics*, 176(1):611–623, 2007.

C.H. Kao, Z.B. Zeng, and R.D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203, 1999.

CM Kendziorski, M. Chen, M. Yuan, H. Lan, and AD Attie. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, 62(1):19–27, 2006.

M.K. Kerr, M. Martin, and G.A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

E.S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185, 1989.

T. Lassmann, O. Frings, and E.L.L. Sonnhammer. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research*, 2008.

C. Li and W.H. Wong. 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(1):31–36, 2001.

D.J. Lockhart, E.L. Brown, G.G. Wong, M. Chee, and T.R. Gingeras. Expression monitoring by hybridization to high density oligonucleotide arrays. *Nature Biotechnol*, 14(13):1675–1680, 1996.

ZW Luo, E. Potokina, A. Druka, R. Wise, R. Waugh, and MJ Kearsey. SFP genotyping from Affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics*, 176(2):789–800, 2007.

M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509, 2008.

M. McGee and Z. Chen. New spiked-in probe sets for the Affymetrix HGU-133A latin square experiment. *COBRA preprint Series*, 5, 2006.

F. Mosteller and J.W. Tukey. Data analysis and regression. A second course in statistics. 1977.

G.C.W.M. Mulders, G.T. Barkema, and E. Carlon. Inverse Langmuir method for oligonucleotide microarray analysis. *BMC bioinformatics*, 10(1):64, 2009.

F. Naef and M.O. Magnasco. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68(1):11906.

M.A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

N. Ono, S. Suzuki, C. Furusawa, T. Agata, A. Kashiwagi, H. Shimizu, and T. Yomo. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics*, 24(10):1278–1285, 2008.

W. Pan. A comparative review of statistical methods for discovering differentially ex-

pressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.

M.D. Robinson and G.K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881, 2007.

J. Ronald, J.M. Akey, J. Whittle, E.N. Smith, G. Yvert, and L. Kruglyak. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Research*, 15(2):284, 2005.

J. Rougemont, A. Amzallag, C. Iseli, L. Farinelli, I. Xenarios, and F. Naef. Probabilistic base calling of Solexa sequencing data. *BMC bioinformatics*, 9(1):431, 2008.

J. SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the national academy of sciences of the United States of America*, 95(4):1460, 1998.

J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728, 2005.

S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature genetics*, 22:281–285, 1999.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 (9):5116, 2001.

C. Welinan. *Inverse kinematics and geometric constraints for articulated figure manipulation.* PhD thesis, Simon Fraser University, 1993.

MH Wilkins, AR Stokes, and HR Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738, 1953.

R. Wolfinger and M. Oconnell. Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3):233–243, 1993.

R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

Z. Wu and R.A. Irizarry. A statistical framework for the analysis of microarray probe-level data. *Ann. Appl. Stat*, 1:333–357, 2007.

Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

S. Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163 (2):789, 2003.

S. Xu. An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity*, 2010.

Z.B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457, 1994.

H. Zhan, X. Chen, and S. Xu. A Stochastic Expectation and Maximization (SEM) Algorithm for Detecting Quantitative Trait Associated Genes. *Bioinformatics*, page submitted, 2010.

L. Zhang, M.F. Miles, and K.D. Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nature biotechnology*, 21(7):818–821, 2003.

C. Zhongxue, M.G. Monnie, L. Qingzhong, K. Megan, D. Youping, and S. Richard. A distribution-free convolution model for background correction of oligonucleotide microarray data. *BMC Genomics*, 10(Suppl):S19.