Bayesian Model-based Methods for the Analysis of DNA Microarrays with Survival, Genetic, and Sequence Data

Jonathan A. L. Gelfond

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2007

Approved by:

Advisor: Joseph G. Ibrahim
Co-advisor: Fei Zou
Co-advisor: Mayetri Gupta
Reader: Fred A. Wright
Reader: David Threadgill

ABSTRACT

JONATHAN GELFOND:
Bayesian Model-based Methods for the Analysis of DNA Microarrays with Survival, Genetic, and Sequence Data
(Under the direction of Dr. Joseph G. Ibrahim)

DNA microarrays measure the expression of thousands of genes or DNA fragments simultaneously in which probes have specific complementary hybridization. Gene expression or microarray data analysis problems have a prominent role in the biostatistics, biological sciences, and clinical medicine. The first paper proposes a method for finding associations between the survival time of the subjects and the gene expression of tumor microarrays. Measurement error is known to bias the estimates for survival regression coefficients, and this method minimizes bias. The latent variable model is shown to detect associations between potentially important genes and survival in a breast cancer dataset that conventional models did not detect, and the method is demonstrated to have robustness to misspecification with simulated data. The second paper considers the Expression Quantitative Trait Loci (eQTL) detection problem. An eQTL is a genetic locus that influences gene expression, and the major challenges with this type of data are multiple testing and computational issues. The proposed method extends the Mixture Over Marker (MOM) model to include a structured prior probability that accounts for

the transcript location. The new technique exploits the fact that genetic markers are more likely to influence transcripts that share the same location on the genome. The third paper improves the analysis of Chromatin (Ch)-Immunoprecipitation (IP) (ChIP) microarray data. ChIP-chip data analysis estimates the motif of specific Transcription Factor Binding Sites (TFBSs) by comparing the IP DNA sample that is enriched for the TFBS and a control sample of general genomic DNA. The probes on the ChIP-chip array are uniformly spaced on the genome, and the probes that have relatively high intensity in the IP sample will have corresponding sequences that are likely to contain the TFBS motif. Present analytical methods use the array data to discover peaks or regions of IP enrichment then analyze the sequences of these peaks in a separate procedure to discover the motif. The proposed model will integrate enrichment peak finding and motif discovery through a Hidden Markov Model (HMM). Performance comparisons are made between the proposed HMM and the previously developed methods.

ACKNOWLEDGMENTS

I would especially like to thank Dr. Joseph Ibrahim for his mentorship and his research advice during the completion of this dissertation. Also, I would like to thank Fei Zou and Mayetri Gupta for their important contributions and guidance as well as the other members of my committee, Fred Wright and David Threadgill. I could not have undertaken this dissertation without the generous support of the Howard Hughes Medical Institute Predoctoral Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

cDNA	Complementary DNA
ChIP	Chromatin Immunoprecipitation
DNA	Deoxyribonucleic Acid
EM	Expectation Maximization
FDR	False Discovery Rate
HMM	Hidden Markov Model
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate/Estimator
mRNA	Messenger RNA
SNP	Single Nucleotide Polymorphism
TFBS	Transcription Factor Binding Site

1 Introduction and Literature Review

Microarrays are quantitative assays that can measure the gene expression levels of thousands of transcripts or millions of DNA fragments simultaneously. Since the mid 1990s, this technology has provided a wealth of new biological and medical insights from the discovery of the often subtle influences that experimental conditions can have on gene expression to the recognition of previously unknown cancer subtypes. In the future, one may expect that microarrays or similar technologies will provide insights into gene networks, and that gene expression analysis will be used to guide clinical decisions on chemotherapy to find optimal treatments and avoid unnecessary side effects.

The scientific potential of microarrays is enormous, and the statistical challenges of the technology are nontrivial. First, there is the problem of multiple testing. Thousands or even millions of dependent hypotheses can be tested in a single experiment. The simple p-values and Bonferroni corrections have been enhanced by estimates of the false discovery rates as a means of characterizing the certainties of inferences. Second, these hypotheses are not made independently of one another. The genes do not work independently, and the expression measurements of the genes share parameters between them that should be modeled in order to utilize all of the information on the array. Third, microarrays are indirect measurements that often have several components per transcript. An estimate of the true expression level should be obtained that summarizes these components; these estimates are referred to as the Gene Expression Indices (GEIs). There are many other

difficulties such as normalization for experimental comparisons and outcome prediction based on high dimensional data.

The dissertation is organized as follows. In Chapter 1, the scientific and statistical fundamentals of microarrays are introduced. The first paper develops a measurement error model for time-to-event data and tumor microarrays and is discussed in Chapter 2. Chapter 3 presents the second paper and the analysis of genetics and microarrays. The paper capitalizes on the relationship between genomic location and the genetic control of transcription. The third paper is discussed in Chapter 4, and it concerns statistical methods that use microarrays to discover transcription factor binding sites.

1.1 Fundamentals of Microarrays

Some biological and technological knowledge of how microarrays work is necessary for the development and understanding of analytical methods. The biology that underlies expression microarrays is often referred to as the "Central Dogma of Molecular Biology." For a review see Watson et al. (2004) or Stryer (1995). This is the principle that the information coded in the nucleotide sequence of DNA is *transcribed* into mRNA which is moved to the cytoplasm and *translated* into polypeptides that modulate and enzymatically promote most of the biochemical reactions in a cell. Gene expression is the process of regulated transcription which is vital for cellular differentiation and function. Microarrays are simultaneous measurements of this transcription process of thousands of genes in a tissue or cell culture. Basically, a microarray is a snapshot quantification of how much particular genes are being transcribed.

There are some specifics of polynucleotide molecules like DNA and RNA that are vital to the self-reproductive properties of cells and to microarrays. DNA molecules are sequences of nucleotide bases that are the purines adenosine and guanine and the pyrimidines thymine and cytosine. In RNA, the role of the purine thymine is replaced by uracil. The polynucleotide chains of DNA and RNA will bind according to the hydrogen bonds of their base sequence. The pyrimidines thymine, uracil have corresponding hydrogen bonds with the purine adenine, and cytosine has corresponding hydrogen bonds with guanine. These favorable configurations result in the *complementary* binding of adenine with cytosine and uracil and the binding of guanine with cytosine. Polynucleotide molecules will preferentially bind to other nucleotides that have the complementary sequence, and this process of one nucleotide binding to another is called *hybridization*. There are many varieties of microarrays, but they all depend on the principle of specific hybridization. The various mRNA of the cells are extracted through a technical process, but their sequences remain specific to the gene from which they were transcribed in the cell. The cellular extraction or sample of all of these mRNA is then labeled with either a fluorescent dye or radioactive isotope that binds in a non-specific manner to the mRNA. On the microarray slide is a spot or probe that contains the complementary sequence to a particular gene, say "G1". Sometimes multiple spots will represent the same gene, and these collection of spots are called probe sets. When the sample makes contact with the slide, only the "G1" mRNA will bind to the "G1" probes or spots because of the specific hybridization. There may be tens of thousands of different probe sets each representing different genes on an array. The fluorescent intensity or radioactivity of all of the probe sets can be quantifiably measured by imaging. The resulting image of the slide is segmented into the

different probes, and a summary statistic of the intensity of light or radiation for each probe is obtained by image analysis (Yang et al., 2001).

Two major classes of microarrays are the two color microarrays and the Affymetrix oligonucleotide arrays. Some of the earliest microarrays used a two dye system to obtain a relative quantification of mRNA (Cho et al., 1997). Two different samples of mRNA are labeled with two different fluorescent dyes (e.g. red and green) and are hybridized to the same array. The array probes are spots containing nucleotide sequences complementary to a specific gene. The ratio of red/green of the probe's fluorescent intensity is taken to be a relative measure of the mRNA levels corresponding to the biological states of the two samples. Affymetrix high-density oligonucleotide arrays are synthesized by a proprietary photolithography process that allows the synthesis of up to 10^5 different probes on the same array (Fodor et al., 1993; Lockhart et al., 1996). The probes consist of short complementary sequences of length 25, and the probes are in perfect match/mismatch pairs. The 13^{th} nucleotide in a mismatch probe is not complementary to the transcript sequence whereas the perfect match probes are entirely complementary to the corresponding sequence. A probe set representing a gene will be about 10-20 pairs, each complementary to a different part of the gene's sequence.

1.2 Gene Expression Index

The Gene Expression Index (GEI) is the scalar valued summary statistic of the gene expression level based on the probe set data. For example, a probe in a two dye system has red and green intensity components that often correspond to a control sample (green)

and an experimental sample (red) (Pollack et al., 2002). The log-ratio of red/green summarizes the two measurements by giving an estimate of the log of the ratio of the concentrations in the two biological states. The motivation for using the log-ratio is manifold. The ratio gives an easily interpretable comparison of the two concentrations and may reduce variation from multiplicative noise that might be present in the measurement variability of both the red and the green channels (Ideker et al., 2000; Rocke and Durbin, 2001). Taking the log acts to symmetrize the distribution of the ratio; the distribution of the ratio is strictly positive and positively skewed. The log-ratio is not universal though, sometimes the red and the green components are analyzed separately (Wolfinger et al., 2001; Kerr et al., 2002).

In Affymetrix arrays, several different probes measure the presence of different components of the gene's sequence, and the models for the GEI are more complex. Early studies used the average difference model (Lipshutz et al., 1999) as the GEI. Let P_{gij}^{PM} be the intensity measurement of the g^{th} gene's i^{th} measurement of the j^{th} perfect match probe where $j = 1, \ldots, J$, and let P_{gij}^{MM} be the corresponding mismatch probe. The statistical model has the form

$$P_{gij}^{PM} = \nu_{gj} + \theta_{gi} + \epsilon_{gij}^{PM} \quad (1)$$

$$P_{gij}^{MM} = \nu_{gj} + \epsilon_{gij}^{MM} \quad (2)$$

where ν_{gj} is the common background effect, θ_i is the gene expression effect, and ϵ_{ij} is the random error. The average difference GEI is given by

$$AD_{gi} = \frac{1}{J} \sum_{j=1}^{J} P_{gij}^{PM} - P_{gij}^{MM}. \quad (3)$$

The motivation for the average difference model is the elimination of ν background variability, and the utilization of all probe data. However, there are some problems with this model. It was seen that θ_{gi} could sometimes be negative which is problematic for estimates proportional to concentration. Also, the average difference model did not take into account the reduced coefficient of variation for higher values. Li and Wong (2001) proposed a model that had parameters that reflected the various probe sensitivities to their target. The model is as follows

$$P_{gij}^{PM} = \nu_{gj} + \alpha_{gj}\theta_{gi} + \phi_{gj}\theta_{gi} + \epsilon_{gij}^{PM} \quad (4)$$

$$P_{gij}^{MM} = \nu_{gj} + \alpha_{gj}\theta_{gi} + \epsilon_{gij}^{MM} \quad (5)$$

where ϕ_{gi} are the specific probe sensitivities, and α_{gi} represent the probe sensitivities to non-specific binding. The ϕ_{gj} parameter had the constraint $\sum_{j=1}^{J} \phi_{gj}^2 = J$. Nevertheless, Li and Wong showed that their model could give improved GEIs that would more accurately detect differential expression in different biological states. Since the Li and Wong model, there have been many competing models that give GEIs for Affymetrix data such as Robust Microarray Analysis (Irizarry et al., 2003) and a mixed model approach of Hsieh et al. (2003). Both of these methods use a transformation of the intensity measurements.

1.3 Microarrays and Multiple Testing

The types of hypotheses most common are those that test for the existence of an association between a gene's expression and the biological states of the collected sample. For example, experiments have found associations between gene expression and cell cycle

Table 1: **Possible outcomes in testing m hypotheses**

	Declared non-significant	Declared significant	Total
True Null	U	V	m_0
True Alternative	T	S	$m - m_0$
	$m - R$	R	m

(Cho et al., 1997; MacAlpine and Bell, 2005), irradiation exposure (Tusher et al., 2001; Snyder and Morgan, 2004; Burns and El-Deiry, 2003), exposure to various compounds (Bartosiewicz et al., 2001; Lobenhofer et al., 2004; Shultz et al., 2001), and survival times (Hastie et al., 2000; Beer et al., 2002; Vijver et al., 2002). In these studies, there are thousands of hypotheses because there are thousands of genes. We will refer to these hypotheses as H_i where i represents the gene identification number, and $H_i = 0$ when the gene is under the null (no association), and $H_i = 1$ when the gene is under the alternative (association). Classical methods generally focused on the development of testing procedures that controlled the type I error rate for one or a few tests of hypotheses (Lehmann, 2005). New testing procedures were developed to increase the power of inferences in this setting. Benjamini and Hochberg (1995) introduced the False Discovery Rate (FDR) approach with the following Table 1.

The quantity $\frac{m_0}{m}$ is the proportion of hypotheses that are truly null, and it is often called π_0. The family-wise error rate (FWER) is defined to be

$$\text{FWER} = P\{V > 0\} = E\{I[V > 0]\}. \tag{6}$$

FWER is the probability that there was at least one rejection of a null hypothesis or false discovery. If the truth concerning the m hypotheses was known such that the table could be constructed then the FDR would be given by

$$\text{FDR} = I[R > 0]\frac{V}{R} = I[R > 0]\frac{V}{V + S}. \tag{7}$$

Since the truth concerning these hypotheses is not known, the FDR was defined to be

$$\text{FDR} = E\left[\frac{V}{R}|R > 0\right] P\{R > 0\} \tag{8}$$

where $\frac{V}{R} = 0$ when $R = 0$. There is a closely rated concept of the positive false discovery rate pFDR $= E[\frac{V}{R}|R > 0]$, but for microarray analyses $P\{R > 0\} \approx 1$ so that FDR $\approx$ pFDR. A simple interpretation of the FDR is the expected proportion of false discoveries in the set that is rejected. As Ge et al. (2003) point out,

$$\text{FDR} \leq \text{pFDR} \leq \text{FWER}. \tag{9}$$

The above inequality is readily shown by writing the quantities as expectations and illustrates the utility of the FDR to avoid the excessive strictness of controlling the FWER. Biologists and clinicians are less concerned about the probability of making a single false discovery (FWER) than in finding a set of genes that has a FDR that is controlled in some sense.

The FWER, FDR or pFDR may be controlled in three ways, and these three ways depend on the conditions under which these expectations are estimated. *Strong* control is achieved when these expectations are bounded from above conditional on any set of null hypotheses being true. *Exact* control is a bounding of the expectations under the condition of knowing the truth concerning the null hypotheses. *Weak* control is satisfied

when the expectations are controlled under the *complete null* hypothesis which states that all hypotheses are under the null. An important early result that Benjamini and Hochberg (1995) proved for independent null hypotheses is that the FDR is controlled in the strong sense under the following procedure. Order the p-values and let these be $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$. If one defines k to be the largest i such that $P_{(i)} \leq \frac{i}{m}q^*$ and rejects the $H_{(r)}$ for $r \in \{1, \ldots, k\}$, then the FDR will be less than $\pi_0 q^*$. They do not estimate π_0 so that they simply bound the FDR by q^*. Benjamini and Yekutieli (2001) extended this type of procedure to handle arbitrary dependence in the test statistics. This procedure is as follows. If one rejects k hypotheses when k is the largest i for which $P_{(i)} \leq \frac{i}{\sum_{l=1}^{m} 1/l}q$, then the FDR is no greater than $\pi_0 q$.

The Benjamini Hochberg (BH) procedure is an example of a *stepwise* procedure. A stepwise procedure is one that involves using the rejection decision of other tests to influence the rejection of another. Specifically, the BH procedure is a *step-up* procedure that starts at the least significant test (i.e. the largest p-value). A *step-down* test such as the Westfall and Young (1993) procedure to control the FWER starts at the most significant test with the smallest p-value. These stepwise procedures are contrasted with the *single-step* procedures like the Bonferroni, Sidak, minP and maxT p-value adjustments (Ge et al., 2003). These are called single-step because the rejection decision of one test does not involve the decisions concerning other tests.

Efron et al. (2001) and Storey (2003) utilize the connection between the FDR and a Bayesian interpretation of the multiple testing problem. Efron et al. (2001) use an empirical Bayes method to estimate the *local false discovery rate* that is the posterior probability of a hypothesis being under the null given that it is rejected. The discussion

of the connection between the posterior probability and the FDR will be continued later. The pFDR may be written as Storey (2002) did

$$\text{pFDR}(p) = P\{H_i = 0 | p_i < p\} = \frac{\pi_0 P\{p_i < p | H_i = 0\}}{P\{p_i < p)\}} \tag{10}$$

with the exception that the p-value has been substituted for the test statistic. This Bayesian construction leaves the problem of estimating the value of π_0 which Storey does by examining the distribution of the p-values. Storey argues that in an experiment with $\pi_0 > 0$, the density of p-values over the interval $[0, 1]$ should become flat over some subinterval $[\lambda, 1]$. Given a choice of λ, an estimate of π_0 is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m}. \tag{11}$$

This estimate of π_0 is conservative ($\hat{\pi}_0(\lambda) > \pi_0$) because some p_i under the alternative could be greater than λ. The number of rejections $R(p)$ is taken to be a function of the p-value cutoff. He then estimates the pFDR as

$$\widehat{\text{pFDR}}_\lambda(p) = \frac{\hat{\pi}_0(\lambda) p}{\widehat{Pr}(P \leq p)[1 - (1-p)^m]}. \tag{12}$$

In the above equation, the numerator $\hat{\pi}_0(\lambda)p$ is an estimate of the probability of the false positive, and the denominator is the product of estimates for the probability of rejection given p ($\widehat{Pr}(P \leq p)$) and the probability that $R(p) > 0$ ($[1 - (1-p)^m]$). The probability for rejection $\widehat{Pr}(P \leq p)$ is estimated by the observed rejection rate $R(p)/m$, and the probability that $R(p) > 0$ is conservatively estimated under that case that the null hypotheses are independent. Storey et al. (2004) demonstrates that his method of pFDR estimation provides strong control when the test statistics are independent or exhibit weak dependence. Storey et al. (2004) defines weak dependence to be the condition of

$V(p)/m_0$ and $S(t)/(m - m_0)$ converging almost surely as $m \to \infty$. Weak dependence holds for dependence in finite blocks and some other special cases, but it is not clear that the correlation in expression data exhibits weak dependence. Specifically, there are only a finite number of genes so that the meaning of asymptotic and continuity in the empirical distribution cannot directly be applied.

Resampling methods like those of Yekutieli and Benjamini (1999) (YB) provide FDR control under general dependence structure. Yekutieli and Benjamini (1999) use resampling to control FDR in a similar manner as the Westfall and Young (1993) procedure for controlling the FWER. Yekutieli correctly notes the FWER estimation through resampling only requires that the null hypotheses rejected V, but FDR estimation through resampling requires that the number of true alternatives (S) is estimated giving

$$\text{FDR}^{\text{est}}(p) = E_{V^*(p)}\left[\frac{V^*(p)}{V^*(p) + \hat{s}(p)}\right] \tag{13}$$

where p is the p-value threshold for significance. The YB procedure has an estimate of $S = \hat{s}$ that is negatively biased to ensure that the estimate of the FDR is conservative. To this end, YB suggested using $\hat{s} = mp$. Reiner et al. (2003) compared the performance in terms of power of the YB resampling method and the BH procedure. They concluded that the YB resampling method provided small increases in power over the BH procedure.

Storey's idea of estimating π_0 based on the density of the p-values is similar in spirit to the Beta-Uniform Mixture (BUM) method of Pounds and Morris (2003). In this method, the density of the p-values is modeled by a BUM given by

$$f(p) = \lambda + (1 - \lambda)ap^{a-1} \tag{14}$$

for $p \in [0, 1]$. They argue that an upper bound and thus a conservative estimate for π_0 is

given by $\hat{\pi_0} \approx \hat{\lambda} + (1 - \hat{\lambda})\hat{a}$. The density of p-values f is written as a mixture distribution with uniform component $\hat{\pi_0} I[p \in [0, 1]]$ corresponding to the p-values under the null which have cumulative distribution function $F_0(p) = p$, and the alternative component

$$\frac{f(p) - \hat{\pi_0}}{1 - \hat{\pi_0}} \tag{15}$$

which has cumulative density $F_a(p)$. For any p-value cutoff τ, an upper bound of the FDR is given by

$$\widehat{\text{FDR}}_{ub} = \frac{\hat{\pi}_0 F_0(\tau)}{\hat{\pi}_0 F_0(\tau) + (1 - \hat{\pi}_0) F_a(\tau)}. \tag{16}$$

However, the BUM method does not model any dependence structure or address the dependence theoretically. Broberg (2005) discusses the performance of the BUM method as well as other methods under dependence, and found that the BUM method performs reasonably well, but as dependence increases, the BUM method, like other methods, has worsening performance.

There are a number of methods for estimating the FDR directly in terms of a posterior probability. Efron et al. (2001) originally proposed the equivalence between local false discovery rates and posterior probability, but Newton et al. (2001) developed a fully Bayesian model for posterior probabilities involving a parametric hierarchical model for two color microarray data. The model estimated the mixing proportion p of genes that were differentially expressed using an Expectation Maximization (EM) (Dempster et al., 1977) algorithm. The complete data involved an unobserved indicator variable z_g that represented whether or not transcript g was differentially expressed. They advocated using the first-order approximation of the posterior odds

$$\text{odds} = \frac{P(z_g = 1|D)}{P(z_g = 0|D)} \approx \frac{p_A(r, g)\hat{p}}{p_0(r, g)(1 - \hat{p})} \tag{17}$$

where $p_A(r, g)$ and $p_0(r, g)$ are the parametric densities of the red (r) and green (g) spot intensities under the alternative (differential expression) and the null respectively. Kendziorski et al. (2003) extended this model to include different parametric assumptions. Kendziorski used the posterior probability in an empirical Bayes framework as a decision rule. These parametric models were extended to a semiparametric error model by Newton et al. (2004a). Newton proposed using the average posterior probability to estimate the FDR and control the FDR. He defined β_g to be the posterior probability of the null hypothesis for gene g. The β_g are then ranked from smallest to largest, and if β_g is less than some κ, then the genes are identified as differentially expressed. One controls the FDR in this framework by this estimate

$$\widehat{\text{FDR}}(\kappa) = \frac{\sum_g \beta_g I[\beta_g \leq \kappa]}{\sum_g I[\beta_g \leq \kappa]} \leq \alpha. \tag{18}$$

Clearly, $\widehat{\text{FDR}}$ can be controlled by choosing an appropriate κ.

Considering the FDR as an *average* of posterior probabilities exposes a potential problem in some FDR procedures like Storey's q-value. This averaging quality of the pFDR has been proved by Efron and Tibshirani (2002). In short, the FDR is problematic because placing a bound on the average (FDR) of a set does not bound the members of a set (posterior probabilities). Liao et al. (2004) point out the differences between the posterior probabilities

$$P\{H_i = 0 | p_i = p\} = \frac{\pi_0}{\pi_0 + (1 - \pi_0) f_a(p)} \tag{19}$$

and Storey's pFDR

$$P\{H_i = 0 | p_i \leq p\} = \frac{\pi_0 p}{\pi_0 p + (1 - \pi_0) F_a(p)} \tag{20}$$

where f_a and F_a are the density and the cdf under the alternative. The right hand side in the above equation matches the FDR_{ub} in the BUM method. The difficulties with the pFDR arise when considering inferences on specific genes. For example, it is possible that the posterior probabilities for being under the null have a highly positively skewed distribution, and the pFDR can be controlled but the gene of least significance could have a posterior probabilities of being under the null equal to 0.99. Glonek and Soloman (2003) give more examples of these poor decisions resulting in blindly controlling the pFDR. This motivates the development of local FDR methods that approximate the posterior probabilities such as Liao et al. (2004) and Efron (2004).

Other methods have been suggested to control the accuracy of multiple inferences in microarray data. Versions of the negative predictive value for detecting differential expression have been used by Liao et al. (2004) and Genovese and Wasserman (2002). Also, Ibrahim et al. (2002) presented a parametric Bayesian model for modeling optimal inferences concerning differential expression. This model includes correlation between the genes as a form of dependency which was induced by the structure of the hyperparameters. The ratio of the mean expression levels of different states for gene g is given by ξ_g. The posterior probability γ_g is defined as $P(\xi_g > 1|D)$, and a threshold $\gamma_0 \in [\frac{1}{2}, 1]$ is then selected such that gene g is differentially expressed if $|\gamma_g - \frac{1}{2}| \leq \gamma_0 - \frac{1}{2}$. The γ_0 threshold could then be set by using the *L measure* criterion. This model selection criterion was developed by Ibrahim and Laud (1994), and it balances the posterior squared predictive error and the posterior variance of the predictions for future observations. The optimal model minimizes the L measure. In the gene expression model, several levels of γ_0 were assessed with the L measure and the optimal γ_0 determined the list of genes declared to

be differentially expressed. This list is approximately optimal in terms of the L measure, and the list is determined without selecting an arbitrary p-value cutoff.

2 Microarrays and Survival Data

2.1 Cancer and Gene Expression

The first paper presents a method for finding associations between time-to-event data and microarrays of tumors. Microarrays have been used to study cancer in many ways. First, scientists saw in microarrays a technique for differentiating cancer types that cannot be differentiated by other means. Microarrays are now capable of measuring every gene in a cell giving a near complete picture of the *transcriptome*. The transcriptome contains vast amount of information about tumors and can be used to differentiate different types of cancer. This process of classification of high-dimensional transcriptomes into distinct subtypes is a statistical problem known as unsupervised classification or clustering (Hastie et al., 2001). Methods of unsupervised classification include hierarchical clustering (Eisen et al., 2001), self-organizing maps (Tamayo et al., 1999), and some more statistical methods like Parmigiani et al. (2002). Microarrays are not simply a taxonomic tool, but the transcriptome gives biological insight as well (Golub et al., 1999). Unsupervised clustering techniques can be used to find clusters of the genes, and biologists recognize that genes within known pathways often are found within these clusters (Golub et al., 1999; Perou et al., 2000). The discoveries of pathways and the gene expression patterns involved in disease is a critical component of finding targets for potential therapies (Evans and Guy, 2004).

The transcriptomes of tumor samples have also been successfully used to predict survival for several different types of cancer including lung adenocarcinoma (Beer et al., 2002), breast cancer (Sorlie et al., 2001; Sotiriou et al., 2003), hepatocellular carcinoma(Lee et al., 2004), and leukemia (Chiaretti et al., 2004). The combination of survival data and expression data has become an increasingly important and common analysis problem. One of the fundamental difficulties in analysis of expression and survival data is that the number of predictors (transcripts) is much larger than the number of independent survival times. This leads to a nonidentifiability problem in estimating regression parameters. Classical Principle Components Regression (PCR) involves using the principle components of the data matrix as the linear predictors (Nguyen and Rocke, 2002). The principle components with the smallest eigenvalues are discarded from analysis thus reducing the dimension of the predictor matrix. However, Nguyen and Rocke (2002) have shown that PCR performs poorly relative to the method of Partial Least Squares (PLS) in predicting tumor classification based on expression profiles, but PLS is not optimal in any reasonable way as shown by Butler and Denham (2000). The poor performance of PCR is not surprising because principle components are an orthogonal decomposition of the total variation of only the predictor matrix, and they are not necessarily associated with the variational patterns correlated with survival. For example, the application of unsupervised clustering techniques to tumor data can lead to classes that are not associated with survival as seen by Bair and Tibshirani (2004).

Several strategies have been developed for predicting survival based on expression data, and most of them used supervised methods of classification. Supervised classification is a technique in which the classes (here the survival data) of the objects (here the

transcriptomes) are known in advance of the model construction (Hastie et al., 2001). This is contrasted to unsupervised clustering methods like those that look for previously unknown classes in the data. Naturally, survival times are continuous and censored in nature, so forming discrete classes is often arbitrary. If these classes are treated as known, then the stochastic properties are ignored which can lead to overfitting (Bair and Tibshirani, 2004). Nevertheless, there exist many mathematical techniques for supervised classification such as neural networks (Wei et al., 2005), support vector machines (Lee et al., 2003), and penalized logistic regression (Shen and Tan, 2005) that have been applied to tumor gene expression. Hastie et al. (2000) developed a "gene shaving" procedure that is related to principle components analysis and takes advantage of the survival times in order to find clusters of genes that are associated with survival. Gene shaving accomplished this by selecting genes into clusters by a balance of both their associations with survival as well as correlations with other genes. There are several tuning parameters that must be predetermined in the gene shaving method including the balance parameter that determines the degree to which survival data influences the principle components in the predictor matrix. Bair and Tibshirani (2004) created a semi-supervised method for predicting survival in which only genes most associated with survival were included in a reduced predictor matrix. The reduced predictor matrix was then decomposed into principle components that are used in a predictive model. It is important to notice that the univariate associations with gene expression play a vital role in some of these supervised methods. In the first paper of the proposal, methods are developed for improving the joint modeling of gene expression and survival that take into account the measurement error.

2.2 Measurement Error Models

The measurement error of microarrays is often not modeled directly by methods that link gene expression and survival. Ideally, one would like to consider the variability due to microarray measurement error when making inferences because microarrays have significant amounts of assay noise (Yang et al., 2002). The presence of noise is obvious in the case of assay replication, but assay noise can be confounded with biological variation depending on experimental design. In the case where the biological states are finite (i.e. treatment and control), there is often biological and technical replication of the states. The assay noise in the presence of biological or technical replication is accounted for by estimating the variance within replicates in the manner of t-statistics (Dudoit et al., 2002). No two tumors constitute the same biological state. Unless the same tumor is assayed more than once, the assay noise will be confounded with biological variation between tumors. The analysis of noise in the absence of either biological or technical replication is not straightforward, but it is of interest to account for the effects of assay noise when dealing with time-to-event data. It is a well known phenomena that failure to account for measurement error in covariates results in asymptotic bias of the estimated effect toward the null (Prentice, 1982; Nakamura, 1992). This gives us motivation to develop a model that includes assay noise and avoids biased inference. Tadesse et al. (2005) have recently shown how inferences concerning microarrays and survival can be affected by not accounting for measurement error in Affymetrix microarrays, and we would like to build a similar model for cDNA microarray data.

The aim of the first paper is to construct a model that accounts for the effects mea-

surement error in cDNA microarray experiments on the assessment of associations between gene expression and time-to-event data. We present a Bayesian hierarchical latent variable model linked with a piecewise constant proportional hazards model for the time-to-event data. The latent variable corresponds to the Gene Expression Index, and the hazard function is conditional upon this latent GEI. The model is shown to have favorable properties such as robustness to misspecification and GEIs that do not explicitly depend on platform specific parameters. Platform specific parameters include the sensitivity of the red probe compared to the green probe and the reference sample. We apply the model to a particular breast cancer experiment that previously demonstrated novel subtypes of breast cancer based on gene expression profiles. The time-to-event of interest is time-to-death due to disease.

Characterizing the association between time-to-event data and gene expression is similar to the differential expression problem because event data constitutes a biological state, although the state is complex in that the state space is censored and infinite. The broader problem of differential expression requires that the gene expression for a particular gene on an array is measured or computed. The aforementioned value of a gene's expression is often referred to as the gene expression index (GEI) (Li and Wong, 2001). The GEI is computed in numerous ways depending on the type of array and the model used. The probes or spots on an individual array have complementary subsequences highly specific to the corresponding gene's RNA in the samples. The proposed model extends the additive error-in-variable survival model for Affymetrix data of Tadesse et al. (2005) to two color microarrays with correlated multiplicative errors. The error model is included within the framework of a piecewise exponential survival model. Robustness

analyses are performed, and the model is applied to a breast cancer study dataset.

2.3 The Data Structure

The data analyzed were obtained from experiments performed on breast cancer samples with similar types of cDNA microarrays. There were a total of 85 microarrays of tumor (78), normal tissue (4), and other tissue (3) from 84 individuals, but clinical information was available from only 77 of the individuals who corresponded to tumors. Of these 77 individuals, 75 had time-to-event data available. This subset of the data is the focus of the paper. There were six batches of microarrays, some arrays having 24k probes and others having 8k probes. A common subset of 7,938 probes were selected. In the green channel, one of three batches of standardized reference were used. The red channel for each array consisted of the 75 tumor samples. It has been reported that the differences in the array type and the batch effect due to reference add some variability to the analysis (Sorlie et al., 2001), but this noise is not considered here. The dataset is available from the Stanford Microarray Database (http://genome-www.stanford.edu/microarray) . The endpoint studied was time to death due to disease in months. Survival times were between 0 and 100 months (mean 35.43, median 30.0). Twenty-six of the 75 patients experienced the event after time 0, A Kaplan-Meier curve of the 75 patients is shown in Figure 1.

2.4 The General Model

The goal is to characterize the association between gene expression of particular genes and time-to-event data. The model proposed will integrate both survival times and a

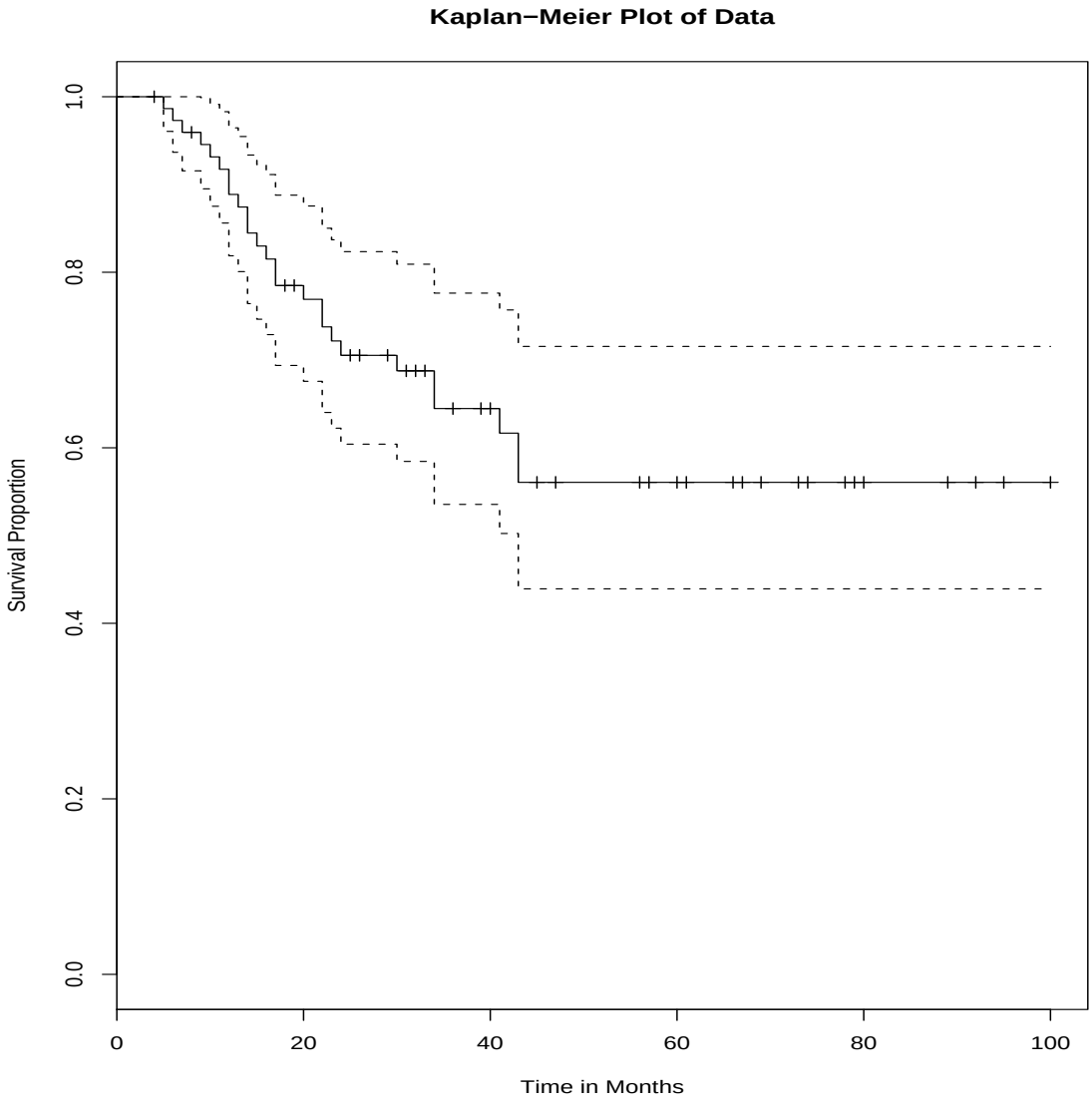

Figure 1: Survival curve for breast cancer dataset

measurement error model. The data is inherently trivariate in that the red and green channels of any probe are potentially correlated with survival and jointly modeled. Some notation will be introduced for this two color data. Each spot on the array will be described as P_{gir} which are vectors of length two whose indices g, i and r refer to the g^{th} gene and the i^{th} individual at the r^{th} replicate respectively. The elements of P_{gir} are R_{gir} and G_{gir} which are the red and green fluorescent measurements of the spot. P_{gir} may be written as $Probe_{gir} \equiv P_{gir} = \begin{bmatrix} R_{gir} \\ G_{gir} \end{bmatrix}$.

The measurement error model is adapted from one proposed by Ideker et al. (2000). Ideker's model consists of a bivariate normal error with an additive component and a multiplicative component. The multiplicative component will be called the spot effect ($spot \equiv S$). The spot effect is the motivation for taking the ratio of R_{gir}/G_{gir}. By dividing R by G, the general assumption is that the multiplicative error will cancel. The additive component is related to the background effect (B). An examination of the data reveals the relationship between the mean probe intensity and the variance of the probe. Figure 2 shows the log of the sample variance plotted against log of the sample mean in the green and red channels of our dataset. There appears to be a strong linear relationship between log(probe mean) and log(probe variance).

Stating the model in equation form we have

$$P_{gir} = M_{gi}S_{gir} + B_{gir}, \tag{21}$$

where

$$S_{gir} \sim N_2\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma^2_{mR} & \rho_m\sigma_{mR}\sigma_{mG} \\ \rho_m\sigma_{mR}\sigma_{mG} & \sigma^2_{mG} \end{bmatrix}\right),$$

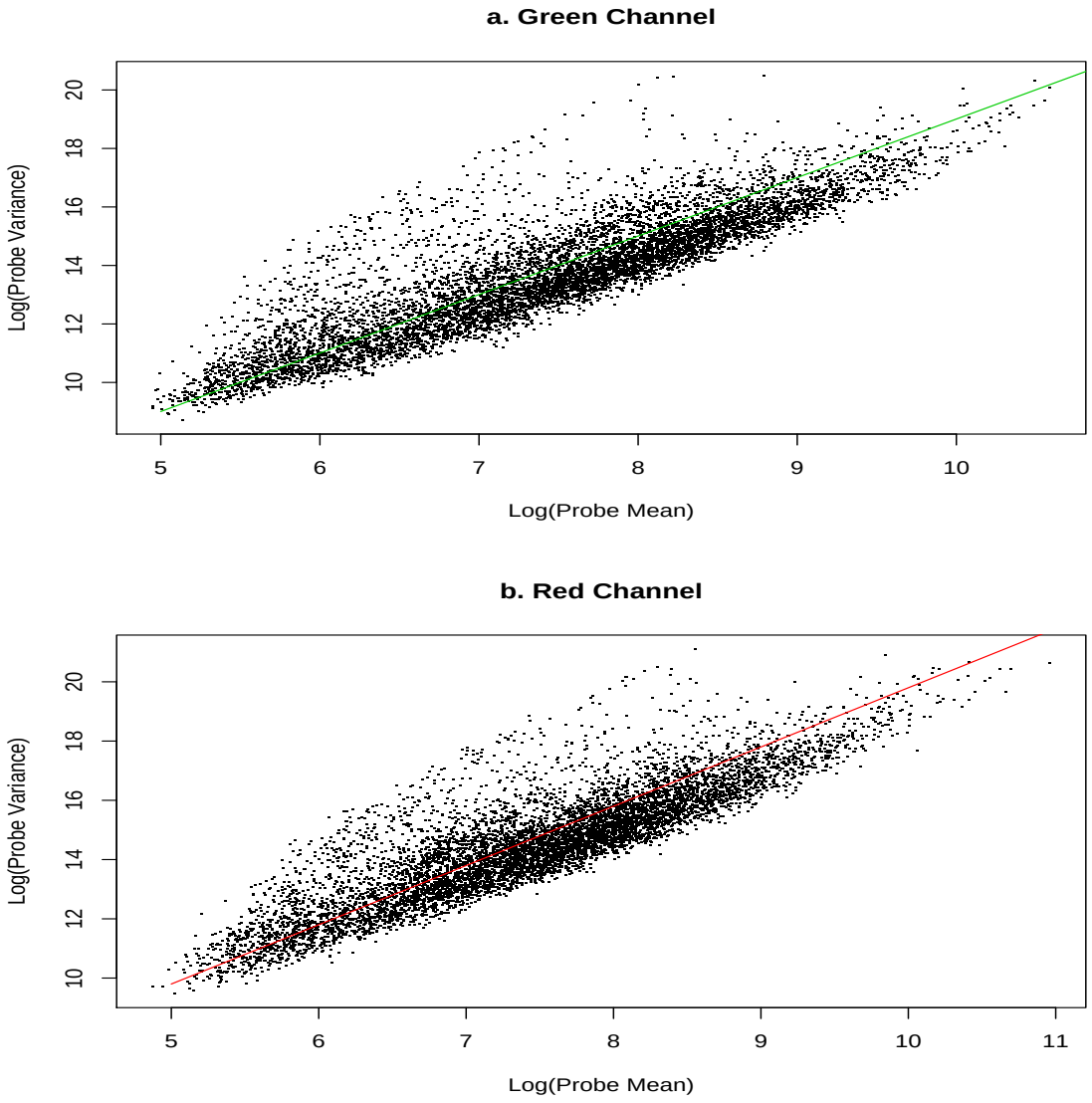

Figure 2: Variance vs Mean Relationship with Model Fit Lines

$$B_{gir} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{aR} & \rho_a \sigma_{aR} \sigma_{aG} \\ \rho_a \sigma_{aR} \sigma_{aG} & \sigma^2_{aG} \end{bmatrix} \right),$$

and

$$M_{gi} = \begin{bmatrix} \mu_{Rgi} & 0 \\ 0 & \mu_{Ggi} \end{bmatrix}.$$

The diagonal elements of M_{gi} are interpreted as the mean intensities for gene g and state i since $E[P_{gir}] = [\mu_{Rgi} \quad \mu_{Ggi}]'$, and this is what motivates the mean vector of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for S_{gir}. The covariance parameters for the multiplicative error are σ^2_{mR}, σ^2_{mG}, and ρ_m which represent the variability due to a multiplicative effect in assay replication of biologically identical samples. Similarly, the covariance parameters for the additive variability due to replication are σ^2_{aR}, σ^2_{aG}, and ρ_a.

There are other models for cDNA data with both additive and multiplicative components. Rocke and Durbin (2001) suggest a log-normal multiplicative error with a normal additive error. This model presents major computational challenges because P_{gir} does not have a standard distribution, and the likelihood cannot be written in closed form. Rocke and Durbin (2001) suggest an iterative fitting procedure on different subsets of genes for the additive and the multiplicative components separately. However, we do not choose this model for three reasons. First, analysis of the residuals of the log transformed data suggest that the log-transformation over-corrects for the relatively small amount of skewness in the data. Second, the difficulty of dealing with a nonstandard distribution adds to an already heavy computational burden. Third, we show in Section 2.6 that the estimation of survival parameters and GEI's with a model based on a normality assump-

tion are robust to this type of misspecification of the multiplicative error distribution. However, even the Ideker model is not identifiable unless there is technical replication in both the red and green channels. In the dataset considered in this paper, we do not have such replication except in a small number of duplicate probes on each array (about 180). An analysis of these probe measurements was performed on the green and the red channels separately, and the estimates of σ^2_{mR} and σ^2_{mG} were found to be approximately equal. With this justification, we set the constraint $\sigma^2_{mR} = \sigma^2_{mG} = \sigma^2_m$ for the purpose of model identifiability. Also, the variance parameters due to the additive components (σ^2_{aR}, σ^2_{aG}, and ρ_a) were found to be very small relative to the multiplicative error, so we set them to zero. The parameters μ_{Rgi} and μ_{Ggi} are the means within a biological state. When a common reference is used, μ_{Ggi} becomes μ_{Gg} and it represents the mean intensity of the reference channel, and μ_{Rgi} is the mean of the sample channel. In experiments with biological replication within a channel, the means of the intensities measured are often considered to be derived from the same underlying population, so that $\mu_{Rgi} = \mu_{Rg}$ for replicates within a biological state. We must account for the biological variability in tumor samples, and thus we consider an additional hierarchical component to the model and take $\mu_{Rgi} = \mu_{Rg}(1 + \beta_{gi})$ The parameter β_{gi} is the latent GEI and represents the i^{th} tumor's and the g^{th} gene's deviation from mean of that gene (μ_{Rg}). β_{gi} is taken to be a truncated normal variable with $\beta_{gi} > -1$ because $\beta_{gi} + 1$ is considered to be proportional to a concentration, and therefore, $\beta_{gi} + 1$ must be positive. The method of identifying the GEI as a latent variable is novel. It is well suited for tumor samples because it gives a structure to the variation in a gene's expression. The structure of the truncated normal distribution acts to resist outlying measurements so that the GEI's have a regression to

the mean. The model can be restated in another equivalent form.

$$R_{gi} = \mu_{Rg}(1+\beta_{gi})\epsilon_{Rgi} \tag{22}$$

$$G_{gi} = \mu_{Gg}\epsilon_{Ggi} \tag{23}$$

where

$$S_{gi} \equiv \begin{bmatrix} \epsilon_{Rgi} \\ \epsilon_{Ggi} \end{bmatrix} \sim \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma_m^2 & \rho_m\sigma_m^2 \\ \rho_m\sigma_m^2 & \sigma_m^2 \end{bmatrix} \right)$$

and

$$\beta_{gi} \sim N_{\{\beta_{gi}>-1\}}(0, \sigma_{bio}^2).$$

We have a simple physical model that assumes that the intensity of a probe (P) is roughly proportional to product of the concentration ([mRNA]) of the target mRNA in the sample and the sensitivity of the probe (ϕ). In equation form we have $P \approx [\text{mRNA}] \times \phi$. The physical model is motivated in part because of the Li and Wong model for Affymetrix data which takes the following form for a single gene:

$$P_{ij} = \nu_j + \theta_i\phi_j + \epsilon_{ij} \tag{24}$$

Here, P_{ij} represents the i^{th} measurement j^{th} probe with sensitivity ϕ_j and background ν_j. θ_i is the gene expression index and ϵ_{ij} is a normally distributed error term. The difference between our model and models like that of Li and Wong is that the GEI of individual i is not a random effect. That is, in the Li and Wong model, the biological variation of GEI's is not modeled explicitly. We extend the form of the Li and Wong model to cDNA data here for the case of a standard reference in the green channel by taking

$$P_{gir} = \Theta_{gi}\Phi_g S_{gir}, \tag{25}$$

where $\Phi_g = \begin{bmatrix} \phi_{Rg} & 0 \\ 0 & \phi_{Gg} \end{bmatrix}$ and $\Theta_{gir} = \begin{bmatrix} [red]_{gi} & 0 \\ 0 & [green]_g \end{bmatrix}$.

The parameters ϕ_{Rg} and ϕ_{Gg} are the platform specific sensitivities of the red and green channels respectively. The Θ_{gi} denotes a matrix whose diagonal elements $[red]_{gi}$ and $[green]_g$ are the concentrations of RNA on the specified array. This model statement is consistent with (21) if we let $\Theta_{gi}\Phi_g = M_{gi}$ and set $B_{gi} = 0$. The problem of gene by dye interaction occurs for some genes when the intensity of the red channel and the green channel respond differently to the same concentration gradient. Using the language of this model, gene by dye interaction can be stated as $\phi_{Rg} \neq \phi_{Gg}$. The connection with this model and the log-ratio can be seen by considering the special case that $\rho_m = 1$. The log-ratio is given by

$$\psi_{gir} \equiv \log(R_{gir}/G_{gir}) = \log\left((\phi_{Rg}/\phi_{Gg})([red]_{gi}/[green]_g)\right). \tag{26}$$

The three deficiencies of ψ_{gir} can be noticed. First, if the values in the red or green channel are negative, then the log-ratio cannot be computed, and this generates missing data despite the clear informativeness of low values. Second, the platform specific parameters of ϕ_{Rg} and ϕ_{Gg} are contained in the GEI. Third, the reference specific parameter μ_{Gg} is also affecting the GEI, and these two problems complicate the interpretation and the cross platform comparisons of the log-ratio. Now, consider the parameter β_{gi}. The parameter can be stated in terms of the ratio of intensity parameters as $\beta_{gi} = (\mu_{Rgi}/\mu_{Rg}) - 1$. According model 29, $\mu_{Rgi} = \phi_{Rg}[red]_{gi}$ and $\mu_{Rg} = \phi_{Rg}[red_g]$ then,

$$\beta_{gi} = (\mu_{Rgi}/\mu_{Rg}) - 1 = ([red_{gi}]/[red]_g) - 1. \tag{27}$$

Thus, β_{gi} does not explicitly depend on platform or reference specific parameters for

the reason that it is a function of the ratio of the mean intensities, and that ratio is not dependent on the probe sensitivity or the reference channel.

The parameter β_{gi} will also be linked to the following piecewise constant hazards survival model. This model divides the survival time axis into J adjacent disjoint intervals $(s_0, s_1], (s_1, s_2], ..., (s_{J-1}, s_J]$ where $0 = s_0 < s_j < s_{j'}$ if $(0 < j < j')$ and $j = 1, \ldots J$. Within each interval is a constant baseline hazard $h_0(y) = \lambda_j$ when $y \in (s_{j-1}, s_j]$. We let $\nu_i = 1$ be the failure indicator for the i^{th} individual ($\nu_i = 0$ otherwise), and let $\delta_{ij} = 1$ if the i^{th} individual was either censored or failed in the j^{th} interval ($\delta_{ij} = 0$ otherwise). The survival component contribution of the likelihood for the i^{th} individual becomes

$$f(y_i|\beta_{gi}, \gamma_c) = \prod_{j=1}^{J} (\lambda_j \exp(\eta_i))^{\delta_{ij}\nu_i} \exp\{-\delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) \right] \exp(\eta_i)\} \tag{28}$$

where $\eta_i = \log(\beta_{gi} + 1)\gamma_g + Z_i'\gamma_c$ is the linear predictor. Z_i is the $p \times 1$ vector of clinical covariates for the i^{th} individual, and γ_c is the corresponding $p \times 1$ vector of coefficients. Note that β_{gi} has been log transformed for comparisons with the log-ratio models.

In this paper, we consider only one gene's ($g = g'$) association with survival at a time so $\beta_{g'}$ refers to the vector of latent GEI's for the $g'^{(th)}$ gene, but P refers to all probe data, that is all of the red and green channel measurements. The model parameters are $\Omega = \{\lambda_j, \beta_{g'i}, \gamma_g', \gamma_{ck}, \mu_{Rg}, \mu_{Gg}, \sigma_m, \rho_m, \sigma_{Bg}\}$.

The dataset consists of $D = \{P_{gi}, Y, \nu_i, \delta_{ij}, \}$. The full likelihood function is the given

by

$$
\begin{aligned}
L(\Omega|D) \propto & \\
& \prod_{i=1}^{n}\prod_{g=1}^{G} \phi_2 \left(P_{gi}; \begin{bmatrix} \mu_{Rg}(1+\beta_{gi}) \\ \mu_{Gg} \end{bmatrix}, \begin{bmatrix} \mu_{Rgi}^2\sigma_m^2 & \rho_m\sigma_m^2\mu_{Rgi}\mu_{Gg} \\ \rho_m\sigma_m^2\mu_{Rgi}\mu_{Gg} & \mu_{Gg}^2\sigma_m^2 \end{bmatrix} \right) \\
\times \quad & \phi_{\{\beta_{gi}>-1\}}(\beta_{gi}; 0, \sigma_{Bg}^2) \\
& \times \prod_{j=1}^{J} \left[\lambda_j e^{(\log(\beta_{g'i}+1)\gamma_{g'}+Z_i\gamma_c)\delta_{ij}\nu_i}\right]^{I[g=g']} \\
\times \quad & \left[\exp\{-\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{k=1}^{j-1}\lambda_k(s_k - s_{k-1})\right] e^{\log(\beta_{g'i}+1)\gamma_{g'}+Z_i\gamma_c}\}\right]^{I[g=g']}. \quad (29)
\end{aligned}
$$

where $\phi_2()$ is the bivariate normal density, and $\phi_{\{\beta_{gi}>-1\}}()$ is the left truncated normal density. Again, $\eta_i = \log(\beta_{g'i}+1)\gamma_{g'} + Z_i\gamma_c$ is the linear predictor for survival involving only one gene (g'). Also, $\mu_{gRi} = \mu_{gR}(1+\beta_{gi})$ for convenience.

The likelihood has two parts. The first part will pertain to the measurement error model, and the second part is the survival model. This dichotomy of the likelihood motivates the two stage fitting procedure described in Section 2.4.2.

2.4.1 Priors

Bayesian models involve the specification of priors as well as the likelihood, therefore the specification of priors will complete our model. We do not have information about parameters from previous studies, and therefore we choose priors that are relatively non-informative or vague. We use the following priors for the parameters

$$\mu_{Rg}^{-1}|\mu_i, \sigma_{\mu_i}^2 \sim N_{\{\mu_{Rg}>0\}}(\mu_i m_{Rg}^{-1}, \sigma_{\mu_i}^2 m_{Rg}^{-2}) \quad [m_{Rg} = \frac{1}{n_g}\sum_{i=1}^{n_g} R_{gi}] \tag{30}$$

$$\mu_{Gg}^{-1}|\mu_i, \sigma_{\mu_i}^2 \sim N_{\{\mu_{Gg}>0\}}(\mu_i m_{Gg}^{-1}, \sigma_{\mu_i}^2 m_{Gg}^{-2}) \quad [m_{Gg} = \frac{1}{n_g}\sum_{i=1}^{n_g} G_{gi}] \tag{31}$$

$$\sigma_m^{-2}|\alpha_m, \omega_m \sim \text{gamma}(\alpha_m, \omega_m) \tag{32}$$

$$\rho_m \sim \text{Unif}(0, 1) \tag{33}$$

$$\sigma_{Bg}|\alpha_B, \omega_B \sim \text{gamma}(\alpha_B, \omega_B) \tag{34}$$

$$\lambda_j|\alpha_0, \omega_0 \sim \text{gamma}(\alpha_0, \omega_0/\lambda_{j-1})(\lambda_0 = 1) \tag{35}$$

$$\gamma_{g'}|\sigma_g^2 \sim N(0, \sigma_{gene}^2) \tag{36}$$

$$\tag{37}$$

The gamma priors on the λ_j's are chosen because they are strictly positive, conjugate, and they induce correlation between adjacent $\lambda's$. Such correlated priors create smoothness in the baseline hazard and were introduced by Arjas and Gasbarra (1994). Such correlated priors are also discussed in Ibrahim et al. (2001). The prior for σ_{Bg} was chosen to be a vague gamma prior; the prior was taken for on σ_{Bg} instead of the precision parameter σ_{Bg}^{-2} because the former is more easily interpreted, and the precision parameter of a truncated normal does not have a conjugate gamma prior. The prior for σ_m^{-2} is a vague gamma prior because this is the conjugate form. A vague normal prior was selected for the survival coefficients γ_g and to let the likelihood drive the inference and make the survival parameters comparable to the Cox model for comparison. The μ parameters in both models had priors that cover the range of the measurements, and a vague prior is placed on μ_{Gg}^{-1} and μ_{Rg}^{-1} instead of the reciprocal to take advantage of

the log-concave posterior which facilitates a more efficient Gibbs sampling scheme. See the appendix for computational details. The array data is scaled to avoid numerical problems. This scaling by m_{Gg} and m_{Rg} results in the choice of $\mu_i = 1$.

2.4.2 Model Fit

Our goal is to fit the model (29) on a gene by gene basis in a computationally efficient manner, and the parameter of interest is $\gamma_{g'}$ because $\gamma_{g'}$ determines the association between gene expression and time-to-event. We could fit the full model likelihood for each gene, but doing so would be computationally expensive because parameters such as $(\beta_{gi}, \mu_{gR}, \text{and} \mu_{gG})$ relating to other genes would then be estimated as well. The number of these nuisance parameters is on the order of $n * G \approx 100,000$. To facilitate a more feasible fitting scheme, the model was fit using an MCMC method in two stages. These two stages correspond to the two parts of the likelihood. In the full likelihood, the first part contains information about the measurement error parameters $(\sigma_m, \rho_m, \sigma_{Bg})$ for all genes, and the second part contains the parameters of the survival model. One may notice that the measurement error parameters are shared across all genes and that one individual gene's contribution to the likelihood should be relatively small. Further, our analysis has shown that these parameters can be estimated to a reasonably high precision by using a large number of genes (≥ 500). Thus, in the first stage of the model fitting, we will estimate the measurement error parameters using likelihood

$$\begin{aligned} Ł(\Omega|D) &= \prod_{i=1}^{n}\prod_{g=1}^{G} \phi_2 \left(P_{gi}; \begin{bmatrix} \mu_{Rg}(1+\beta_{gi}) \\ \mu_{gG} \end{bmatrix}, \begin{bmatrix} \mu_{Rgi}^2\sigma_m^2 & \rho_m\sigma_m^2\mu_{Rgi}\mu_{Gg} \\ \rho_m\sigma_m^2\mu_{Rgi}\mu_{Gg} & \mu_{Gg}^2\sigma_m^2 \end{bmatrix} \right) \\ &\times \phi_{\{\beta_{gi}>-1\}}(\beta_{gi}; 0, \sigma_B^2) \end{aligned} \quad (38)$$

The biological variance parameter σ_B is chosen in this stage to be the same for each gene for computational convenience and to borrow strength across genes. Alternatively, one could select of subset of housekeeping genes thought to have the same low biological variability, and use only these genes to estimate the measurement error parameters. From this model fit, we will use the estimates of the measurement error parameters $\hat{\sigma}_m$ and $\hat{\rho}_m$ and substitute them into (29) and this will constitute the second stage of the model fit:

$$
\begin{aligned}
&L(\Omega|D) \propto \\
&\prod_{i=1}^{n} \phi_2 \left(P_{gi}; \begin{bmatrix} \mu_{Rg}(1+\beta_{gi}) \\ \mu_{g'G} \end{bmatrix}, \begin{bmatrix} \mu_{Rgi}^2 \hat{\sigma}_m^2 & \hat{\rho}_m \hat{\sigma}_m^2 \mu_{Rgi} \mu_{Gg} \\ \hat{\rho}_m \hat{\sigma}_m^2 \mu_{Rgi} \mu_{Gg} & \mu_{Gg}^2 \hat{\sigma}_m^2 \end{bmatrix} \right) \phi_{\{\beta_{g'i} > -1\}}(\beta_{g'i}; 0, \sigma_{Bg'}^2) \\
&\times \prod_{j=1}^{J} (\lambda_j \exp(\eta_i))^{\delta_{ij}\nu_i} \exp\{-\delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) \right] \exp(\eta_i)\}. \qquad (39)
\end{aligned}
$$

The second stage will be applied to each gene, and the parameters associated with the measurement error $(\hat{\sigma}_m, \hat{\rho}_m)$ remain fixed. Further, we found that the model is weakly identifiable when σ_B becomes large ($\sigma_B > 2$). For large σ_B, the parameters σ_B and μ_{Rg} become confounded. So, for the second stage of the analysis, we fixed $\mu_{Rg} = \frac{1}{n}\sum_{i=1}^{n} R_{gi}$. We found that this constraint only had slight influence on the inferences regarding the parameter of interest (γ_g). In order to classify the genes as either significantly associated with an survival or not, we will use the highest posterior density (HPD) intervals for the γ_g parameter. If and only if the interval does not contain 0, then the gene will be included in the list of genes associated with survival.

We fit the models using a Gibbs sampling technique in which samples from the joint posterior $(L(D|\Omega)\pi(\Omega))$ are obtained by successively sampling from the full conditionals for a number of iterations after convergence criteria are met. The log likelihood functions

of the full conditionals for the first stage of the model fit are given below: For notational convenience let

$$
\begin{aligned}
SS_R &= \sum_{g=1}^{G}\sum_{i=1}^{n_g} \frac{(R_{gi} - \mu_{Rg}(1+\beta_{gi}))^2}{(\mu_{Rg}(1+\beta_{gi}))^2} \quad , \\
SS_G &= \sum_{g=1}^{G}\sum_{i=1}^{n_g} \frac{(G_{gi} - \mu_{Gg})^2}{\mu_{Gg}^2} \quad , \\
S_{RG} &= \sum_{g=1}^{G}\sum_{i=1}^{n_g} \frac{(R_{gi} - \mu_{Rg}(1+\beta_{gi}))(G_{gi} - \mu_{Gg})}{\mu_{Gg}\mu_{Rg}(1+\beta_{gi})} \quad , \\
m_{Rg} &= \frac{1}{n_g}\sum_{i=1}^{n_g} R_{gi} \quad , \quad \text{and} \\
m_{Gg} &= \frac{1}{n_g}\sum_{i=1}^{n_g} G_{gi} \quad .
\end{aligned}
$$

We have

$$
\begin{aligned}
p(\mu_{Gg}^{-1}|rest) &\propto \exp\{-\frac{1}{2\sigma_m^2(1-\rho_m^2)}[\frac{1}{\mu_{Gg}^2}\sum_{i=1}^{n_g} G_{gi}^2 \\
&- \frac{2}{\mu_{Gg}}\sum_{i=1}^{n_g}\left(G_{gi}+\rho_m\left(\frac{R_{gi}G_{gi}}{\mu_{Rg}(1+\beta_{gi})}-G_{gi}\right)\right)]\} \\
&\times \mu_{Gg}^{-n_g}\exp\left\{-\frac{(\mu_{Gg}^{-1}-\mu_i m_{Gg}^{-1})^2}{2\sigma_{\mu_i}^2 m_{Gg}^{-2}}\right\} I\left[\mu_{Gg}>0\right] \quad , \\
p(\mu_{Rg}^{-1}|rest) &\propto \exp\{-\frac{1}{2\sigma_m^2(1-\rho_m)}[\frac{1}{\mu_{Rg}^2}\sum_{i=1}^{n_g}\left(\frac{R_{gi}}{1+\beta_{gi}}\right)^2 \\
&- \frac{2}{\mu_{Gg}}\sum_{i=1}^{n_g}\left(\frac{R_{gi}}{1+\beta_{gi}}+\frac{\rho_m}{1+\beta_{gi}}\left(\frac{R_{gi}G_{gi}}{\mu_{Gg}}-R_{gi}\right)\right)]\} \\
&\times {\mu_{Rg}}^{-n_g}\exp\left\{-\frac{(\mu_{Rg}^{-1}-\mu_i m_{Rg}^{-1})^2}{2\sigma_{\mu_i}^2 m_{Rg}^{-2}}\right\} I\left[\mu_{Rg}>0\right] \quad , \\
p(\sigma_m^{-2}|rest) &\sim \text{gamma}\left(\alpha_m+\sum_{g=1}^{G} n_g, \quad \omega_m+\frac{1}{2(1-\rho_m^2)}\left[SS_R+SS_G-2\rho_m S_{RG}\right]\right) \quad , \\
p(\rho_m|rest) &\propto \exp\left\{-\frac{1}{2}\sum_{g=1}^{G} n_g\log(1-\rho_m^2)-\frac{1}{2\sigma_m^2(1-\rho_m^2)}\left(SS_R+SS_G-2\rho_m S_{RG}\right)\right\} \\
&\times I\left[\rho_m\in[0,1]\right] \quad , \\
p(\beta_{gi}|rest) &\propto \exp\{-\frac{1}{2\sigma_m^2(1-\rho_m^2)}[\frac{(R_{gi}-\mu_{Rg}(1+\beta_{gi}))^2}{(\mu_{Rg}(1+\beta_{gi}))^2} \\
&- 2\rho_m\frac{(R_{gi}-\mu_{Rg}(1+\beta_{gi}))(G_{gi}-\mu_{Gg})}{\mu_{Gg}\mu_{Rg}(1+\beta_{gi})}]\} \\
&\times (1+\beta_{gi})^{-1}\exp\left\{-\frac{\beta_{gi}^2}{2\sigma_B^2}\right\} I\left[\beta_{gi}>-1\right] \quad , \text{ and} \\
\sigma_B|rest &\propto (1-\Phi(\frac{-1}{\sigma_B}))^{-\sum_{g=1}^{G} n_G}\sigma_B^{-\sum_{g=1}^{G} n_g+\alpha_B}\exp\left\{-\omega_B-\frac{1}{2\sigma_B^2}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\beta_{gi}^2\right\}.
\end{aligned}
$$

where "$rest$" denotes the data and the remaining parameters.

Computation for the Gibbs sampler was performed using the C language. The full conditionals of ρ_m, σ_m, and β_{gi} were sampled using the Adaptive Rejection with Metropolis Sampling (ARMS) algorithm of Gilks et al. (1995). The μ^{-1} parameters have a log-

concave density, and could be sampled directly using Adaptive Rejection Sampling (ARS) (Gilks and Wild, 1992). The parameter σ_m^{-2} has a gamma distribution which could be sampled using standard statistical algorithms. The ordered overrelaxation technique of Neal (2003) was used when sampling from the σ_m^{-2}, μ_{Rg}^{-1} and μ_{Gg}^{-1} full conditionals to reduce autocorrelation of the Gibbs sampler and improve convergence.

The second stage of the model fit has additional parameters relating to the survival model, and it treats the measurement error parameters σ_m and ρ_m as known by substituting in their estimated values from stage 1. Further, the parameter μ_{Rg} set to m_{Rg} (defined above) for identifiability. Below are the full conditionals for the second stage of the model. For notational convenience, we define Λ_i as the cumulative baseline hazard for individual i

$$\Lambda_i = \sum_{j=1}^{J} \delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{h=1}^{j-1} \lambda_h (s_h - s_{h-1}) \right].$$

We now have

$$
\begin{aligned}
p(\mu_{Gg}^{-1}|rest) &\propto \exp\{-\frac{1}{2\sigma_m^2(1-\rho_m^2)}[\frac{1}{\mu_{Gg}^2}\sum_{i=1}^{n_g} G_{gi}^2 \\
&- \frac{2}{\mu_{Gg}}\sum_{i=1}^{n_g}\left(G_{gi}+\rho_m\left(\frac{R_{gi}G_{gi}}{\mu_{Rg}(1+\beta_{gi})}-G_{gi}\right)\right)]\} \\
&\times \mu_{Gg}^{-n_g}\exp\left\{-\frac{(\mu_{Gg}^{-1}-\mu_0)^2}{2\sigma_{\mu_0}^2}\right\} I\left[\mu_{Gg}>0\right], \\
p(\beta_{gi}|rest) &\propto \exp\{-\frac{1}{2\sigma_m^2(1-\rho_m^2)}[\frac{(R_{gi}-\mu_{Rg}(1+\beta_{gi}))^2}{(\mu_{Rg}(1+\beta_{gi}))^2} \\
&- 2\rho_m\frac{(R_{gi}-\mu_{Rg}(1+\beta_{gi}))(G_{gi}-\mu_{Gg})}{\mu_{Gg}\mu_{Rg}(1+\beta_{gi})}]\} \\
&\times (1+\beta_{gi})^{-1}\exp\left\{-\frac{\beta_{gi}^2}{2\sigma_B^2}\right\} I\left[\beta_{gi}>-1\right] \\
&\times \exp\left\{\nu_i\gamma_g\beta gi-\Lambda_i e^{\log(\beta_{gi}+1)\gamma_{g'}+Z_i'\gamma_c}\right\}, \\
\sigma_B|rest &\propto (1-\Phi(\frac{-1}{\sigma_B}))^{-n_g}\sigma_B^{-n_g+\alpha_B}\exp\left\{-\omega_B-\frac{1}{2\sigma_B^2}\sum_{i=1}^{n_g}\beta_{gi}^2\right\}, \\
p(\gamma_g|rest) &\propto \exp\left[\sum_{i=1}^{n}\gamma_g\nu_i\beta_{gi}-\Lambda_i e^{\log(\beta_{gi}+1)\gamma_{g'}+Z_i'\gamma_c}\right]\times\exp\left[-\frac{1}{\sigma_g^2}\gamma_g\right], \\
p(\gamma_{ck}|rest) &\propto \exp\left[\sum_{i=1}^{n}\gamma_{ck}\nu_i Z_{ik}-\Lambda_i e^{\log(\beta_{gi}+1)\gamma_{g'}+Z_i'\gamma_c}\right]\times\exp\left[-\frac{1}{\sigma_c^2}\gamma_{ck}\right], \quad \text{and} \\
\lambda_j|rest &\sim \text{gamma}\left(\alpha_0+\sum_{i=1}^{n}\nu_i\delta_{ij}, \frac{\omega_o}{\lambda_{j-1}}+\sum_{i=1}^{n}\Delta_{ij}e^{\log(\beta_{gi}+1)\gamma_{g'}+Z_i'\gamma_c}\right)
\end{aligned}
$$

where $\Delta_{ij}=(\min(y_i,s_j)-s_{j-1})^+$.

Again, the ARMS algorithm was used to sample from the posterior distribution within the Gibbs framework for the all of the parameters except λ_j. The γ_{ck} and γ_g parameters have full conditionals that are log-concave so that the ARS algorithm is potentially applicable; however, numerical imprecision sometimes yielded non-concave log-likelihood functions despite the analytical log-concavity of the conditionals. Since ARMS is a more general sampling method, it was used for these parameters. Also, within the ARMS algorithm, the value of the log-likelihood function of the parameters γ_g and β_{gi} was

truncated in the extreme tails to avoid numerical imprecision. Again overrelaxation was used for the μ_{Gg} parameter to improve convergence properties.

2.5 Case Study in Breast Cancer

We use the model in the previous section to examine the breast cancer data described in Section 2.3. As mentioned above, the model was in two stages, measurement error parameter estimation and survival analysis.

2.5.1 Estimating the Measurement Error Parameters

We normalized the microarrays before applying our model. There are many normalization procedures available for cDNA (Yang et al., 2002). However, most of these methods are applied to the log-ratio as opposed to the red and green channel individually. For our purposes, we jointly model the red and green channel instead of modeling $\log(R/G)$. Moreover, there is no replication of samples that is an important component of many normalization procedures. For normalization, we choose to perform a simple scaling procedure as follows. One array without major problems such as poor green or red dye measurements is chosen as the standard, and the red channel measurements from that array are scaled so that the mean of the red channel is equal to the mean of the green channel. Then, all other arrays are scaled so that the means of each channel's probes are equal to the mean of the green channel of the first array. This method was chosen above quantile normalization because it better preserved the correlation between the red and the green channels across arrays. When we compare our method to one that uses the log-ratio, we used the log-ratio normalization procedure used by (Sorlie et al., 2001).

After the arrays were normalized, we estimated the measurement error parameters by sampling 500 probes at random from the original 7,938 probes. Prior parameters were selected as follows: $(\alpha_B, \omega_B) = (2, 0.1)$; $(\alpha_m, \omega_m) = (2, 0.1)$; $(\mu_i, \sigma^2_{\mu_i}) = (1, 100)$. The burn-in period of 10,000 Gibbs sample was used to achieve convergence, and the number of samples used was 50,000. The convergence of the Gibbs sampler was diagnosed with parallel chains by using the Gelman and Rubin $\sqrt{\hat{R}}$ statistic (Gelman and Rubin, 1992). Convergence diagnostics were computed with the R coda package (Plummer et al., 2005). See the Figure 3 for trace plots.

There was some autocorrelation in the parameters that slowed convergence, but the effect on parameter estimation was small as the mean of the posterior estimates were within 1% of their final estimates very early in the chain (i.e. after a few hundred iterations). The measurement error model then yielded estimates for these parameters as follows: $\sigma_B = 0.5752(0.0062)$; $\sigma_m = 0.6082(0.0029)$; $\rho_m = 0.9347(0.0021)$.

These parameters suggest a large amount of variation due to assay noise. The coefficient of variation due to the multiplicative technical error is $\hat{\sigma}_m$, and the correlation of the red and green components of this multiplicative effect is $\hat{\rho}_m$ which suggests that the log-ratio has significant error. These parameters will now be considered fixed in the gene by gene survival analysis stage.

2.5.2 Data Preprocessing

Before survival models are fitted, there is some data preprocessing including gene filtering and imputation of missing data. The large number of genes relative to the number of independent observations makes it beneficial to limit the analysis to a subset of probes

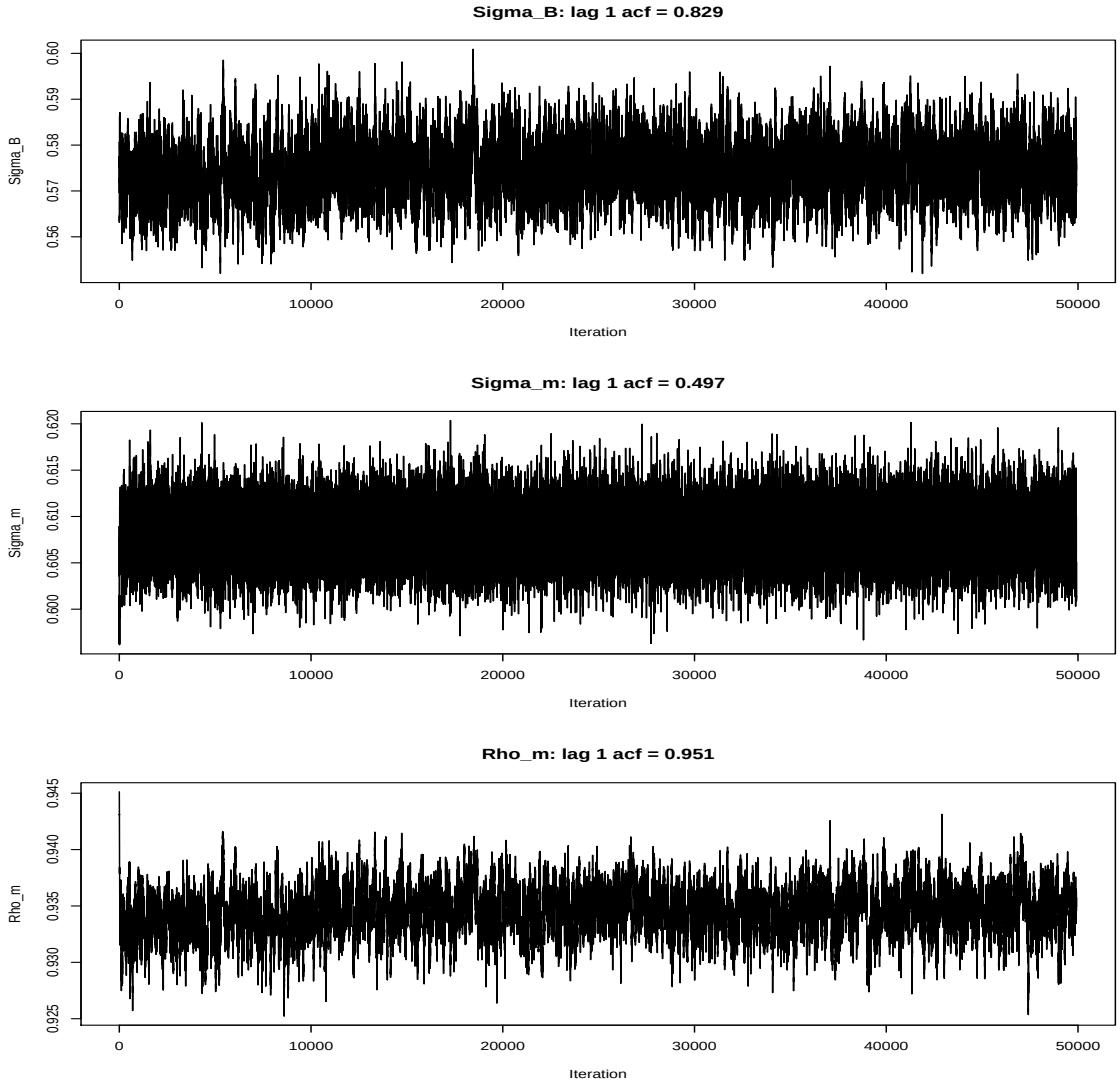

Figure 3: Trace plots of select parameters (σ_B, σ_m, and ρ_m) of measurement error model with lag 1 autocorrelation

that meet some threshold of variability across samples. We used a similar inclusion threshold to that of the original analysis. We considered only probes that had at least 3 samples that were a 4-fold change from the median log-ratio. From that list, we took a subset of those probes which had missing data in the green or red channels for no more than 10 out of the 75 arrays. This left 991 probes for examination, but there were duplicated gene names in the probe list. All duplicate gene names were removed for the survival analysis which left 942 genes. The missing data in the reduced set was then imputed using the log-ratio. Specifically, imputation was performed using the Statistical Analysis for Microarrays package (Tusher et al., 2001) with a K-nearest neighbor algorithm in which K=10.

2.5.3 Results: Genes identified by the Gene Only Model

Our goal is to find a list of genes that are associated with time-to-event in breast cancer. We perform two types of analyses and compare the results with a conventional Cox proportional hazards model. The analysis presented here tests the gene's survival association without additional clinical covariates. For comparison, we fit a Cox proportional hazards model with standard software R Development Core Team (2004a) using the log-ratio as the GEI covariate. When constructing gene lists using the Cox model, the p-value of the corresponding regression parameter was used to determine association. Specifically, lists with genes having a p-value cutoff of < 0.01 for the regression parameter will be compared to lists including genes whose γ_g parameters have 99% HPDs that do not contain 0. The latent variable and the Cox models were fit to the 942 genes. The prior hyperparameters for the survival model are as follows: $(\alpha_B, \omega_B) = (2, 0.1)$; $(\alpha_0, \omega_0) = (0.01, 0.01)$;

$(\mu_i, \sigma^2_{\mu_i}) = (1, 100)$; $\sigma^2_{gene} = 100$.

The Gelman-Rubin statistic was again used to assess convergence, but because of the number of models (942), convergence could not be thoroughly examined except for a few genes. Based on these models, a conservative estimate for the number of iterations needed to achieve convergence was used for all genes. A burn-in of 5,000 cycles, and 10,000 samples were used to summarize the posterior estimates. The results compare lists of genes selected to have a significant association with survival by the by proposed model and the Cox model given in Table 2.

Table 2: Comparison of significant gene lists for Gene Only Model

		Proposed Model		
		Significant	Not Significant	Total
Cox	Significant	65	18	83
Model	Not Significant	13	846	859
	Total	78	864	942

There is significant agreement between the the two lists. For the sake of brevity, we will focus a few important genes. The intersection of the two lists includes genes which have known associations with breast cancer such as the estrogen receptor Perou et al. (2000), gamma glutamyl hydrolase (Rhee et al., 1993), and the angiotensin receptor 1 (AGTR1) gene (De Paepe et al., 2001). Of the 13 genes that were detected by the

Table 3: Clinical Covariates Only Comparison

	Cox Model		Piecewise Exponential	
Covariate	Estimate	(SD)	Estimate	(SD)
Age	0.049	(0.259)	0.014	(0.274)
Tumor Category	0.606	(0.267)	0.640	(0.269)
Grade	0.488	(0.230)	0.483	(0.234)
ER status	0.747	(0.202)	0.726	(0.205)

proposed model only, we have found that some of them have associations with breast cancer such as estrogen regulated LIV-1 protein (Dressman et al., 2001) and the 5T4 oncofetal trophoblast glycoprotein gene (Kopreski et al., 2001).

2.5.4 Results: Inclusion of Clinical Covariates

The clinical covariates of age (< 40), tumor category (1,2,3,4), grade (High, low), and ER status (+/-) were entered into the model. Tumor category corresponds to the size of the tumor, so it was treated as a continuous covariate instead of a factor. First, the clinical covariates were fit without the expression data in order to compare the Cox and the piecewise exponential models. All of these clinical variables were centered and scaled. A burn-in period of 5,000 samples were taken and 10,000 samples were used to compute the posterior estimates. The results in Table 3 show the close agreement between the two models.

The results of the model fit with these covariates and each of the genes are shown in

Table 4. A burn-in of 7,500 cycles and 10,000 samples were used.

Table 4: Comparison of significant gene lists with covariates

		Proposed Model		
		Significant	Not Significant	Total
Cox	Significant	19	11	30
Model	Not Significant	8	904	912
	Total	27	915	942

Many of the 19 genes selected by both models have associations with breast cancer such as Claudin 4 Kominsky et al. (2004). Some of the 8 genes selected only by the latent variable model have associations with breast cancer in the literature such as the somatomedin gene (Byron et al., 2006).

2.6 Robustness Analysis and Operating Characteristics

2.6.1 Deviation from normality in the data

According to the model, the array data in the green channel for a particular probe is normally distributed about the same mean so that $green_{gi} \sim N(\mu_{Gg}, \mu_{Gg}^2 \sigma_m^2)$. One may calculate the scaled residuals in a typical manner of subtracting the sample mean and then dividing by the sample standard deviation for each gene. To examine the validity

of the distributional assumption, we show a histogram of the scaled residuals in Figure 4. The normal density is overlaid.

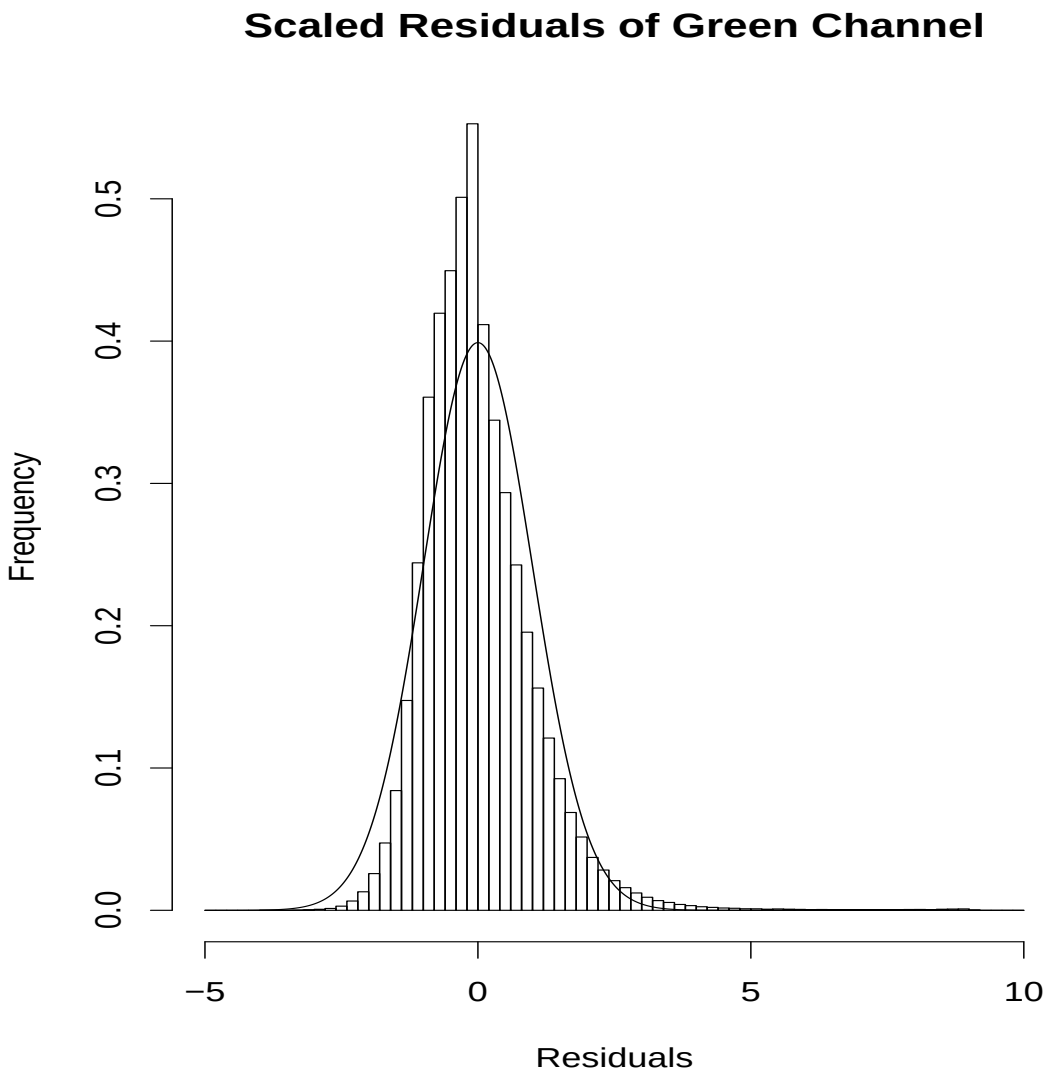

Figure 4: Scaled Residuals with normal density curve.

One can detect that the distribution is skewed to the right with a heavier tail. One could consider a transformation, but transformations dilute the relationship between the mean and the variance. The distribution of the red channel is much more complicated under the model because it is the product of a normal and a truncated normal random

variable.

2.6.2 Simulations demonstrating robustness to nonnormality

In order to characterize the effects of this deviation from normality, we performed a robustness analysis with a simulation. We used the log-normal model of Rocke and Durbin (2001) without the normal additive error to simulate a dataset and applied our two stage model fitting procedure to 200 different datasets with $n = 75$ individuals. The true measurement error parameters of the simulation were $\sigma_m = 0.6$, $\rho_m = 0.9$, and $\sigma_{Bg} = 0.5$. A total of 500 genes were simulated with $\mu_{gR} = |X_{gi}|$ and $\mu_{gG} = \mu_{gR}Y_{gi}$ where $X_{gi} \sim N(10,000, 3,000)$ and $Y_{gi} \sim \text{gamma}(2, 2)$. The estimates (and SD's) from the model fit were ($\hat{\sigma}_m = 0.651 \quad (0.005)$, $\hat{\rho}_m = 0.985 \quad (0.001)$, and $\hat{\sigma}_{Bg} = 0.59 \quad (0.05)$). Then, 200 survival datasets were generated with the survival time y_i being exponentially distributed with rate parameter equal to $\exp[\gamma_g log(\beta_{gi} + 1)]$ with a censoring probability of 0.7. The regression coefficient γ_g was drawn uniformly from the interval $[-2, 2]$, and $\beta_{gi} \sim N_{\{\beta_{gi}>-1\}}(0, 25.0)$. The parameters μ_{Rg} and μ_{Gg} were simulated as above. We are primarily interested in the γ_g parameter, but we also show results of β_{gi}. Figure 5 shows the $\hat{\gamma}_g$ plotted against the true values.

The bars in the plot indicate the 95% HPD intervals. The 95% and 99% HPD intervals contained the true values of γ_g 92.0% and 98.5% of the time respectively, which indicates that the model is estimating γ_g fairly accurately. Also, the $\log(\hat{\beta}_{gi} + 1)$ were highly correlated with the true values, see Figure 6. Another test of robustness of the $\hat{\beta}_{gi}$ as GEI's is the correlation that they have with the conventional log-ratio GEI's. The mean and median correlation of $log(\hat{\beta}_{gi} + 1)$ with the log-ratio estimates for each of the 942

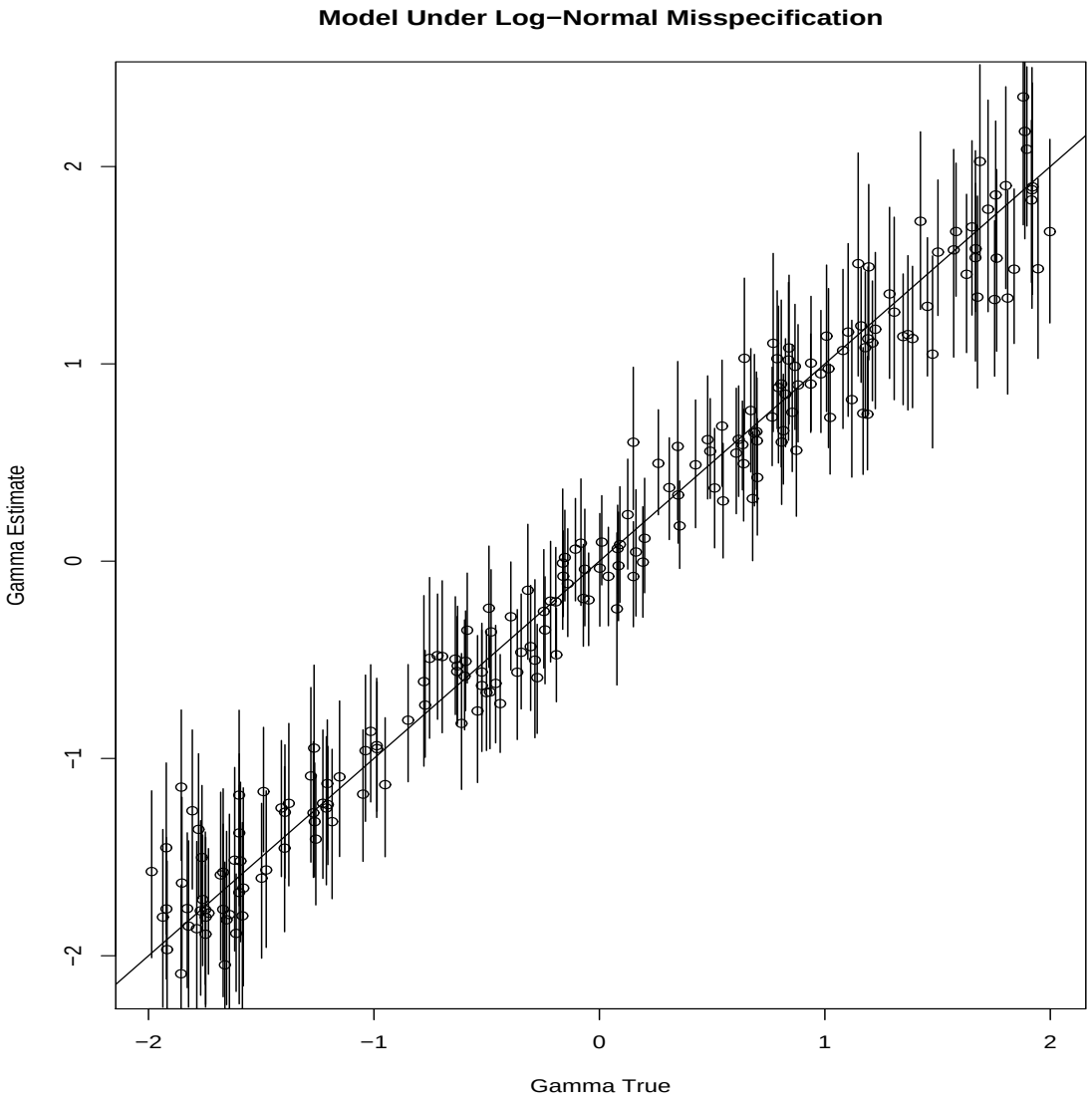

Figure 5: Estimation of Regression Parameter with 95% HPD under misspecified error model.

Table 5: Operating Characteristics Under True Model

N	γ_g	Estimate	(SD)	$\gamma_g \in$ 95%HPD	$\gamma_g \in$ 99%HPD	$0 \notin$ 95%HPD	$0 \notin$ 99%HPD
950	0	0.0004	(0.18)	0.943	0.99	0.057	0.008
25	1	1.11	(0.24)	0.96	1.0	1.0	1.0
25	-1	-1.01	(0.22)	0.96	1.0	1.0	0.96

genes of interest are 0.90 and 0.97 respectively. These high correlations between the log-ratio GEI and the latent GEI estimates suggests substantial agreement of the two estimates of the biological variability present in the data.

An analysis of the operating characteristics of the model demonstrates that the model has good type I and type II error rate control for inference regarding the γ_g parameter. The Ideker et al. (2000) model was used to simulate the datasets, and the same measurement error parameters were used as above with these parameters treated as known. A total of 1,000 datasets with $n = 75$ individuals were simulated with 950 genes under the null ($\gamma_g = 0$) and 50 genes under the alternative ($\gamma_g = 1$ and $\gamma_g = -1$, 25 times each). The results for the simulation are given in Table 5.

Table 5 shows that the properly specified model has no strong evidence of type I error rate inflation and has good power for moderate effect size. Also, the HPDs have accurate coverage probabilities, and the estimated coefficients γ_g show no indication of bias. For comparison, we retested the operating characteristics using the log-normal multiplicative error mentioned above the and using same simulation parameters as well as the same

Table 6: Operating Characteristics under Misspecified Model

N	γ_g	Estimate	(SD)	$\gamma_g \in$ 95%HPD	$\gamma_g \in$ 99%HPD	$0 \notin$ 95%HPD	$0 \notin$ 99%HPD
950	0	0.004	(0.15)	0.941	0.987	0.059	0.013
25	1	0.989	(0.19)	1.0	1.0	1.0	1.0
25	-1	-0.958	(0.16)	0.92	1.0	1.0	1.0

estimated measurement error parameters. Again, 1,000 datasets were simulated with $n = 75$, for fitting the survival model. The results are shown in Table 6.

Table 6 represents the model fit under a grossly misspecified error structure, and this degree of skewness in the error is greater than that of the observed data. Despite this large deviation from normality, the model is seen to be quite robust to this kind of misspecification. One can see a slight inflation of the type I error rate (0.05→0.059 and 0.01→0.013). The power is not seen to decrease, and this may be surprising. However, one must remember that the measurement error is not the same. The estimates of the γ_g coefficients are slightly biased towards the null as would be the case for models that did not account for measurement error, but this bias does not effect the HPD coverage probabilities. Overall, the model shows good type I and type II error rate control under misspecification.

2.7 Discussion

This paper presents a model to find associations between a gene's expression and time-to-event data for cDNA microarrays that accounts for the substantial measurement error. The model for the microarray probes is parametric and creates a GEI which is latent instead using the log-ratio. The model for the time-to-event data is a Bayesian semiparametric piecewise constant hazards model. We fit the model using an MCMC algorithm in a two stage process. The first stage estimates the measurement error parameters, and the second stage uses these estimates in the survival model on a gene by gene basis. A case study with a breast cancer dataset is performed with and without adjusting for clinical covariates. The new model is shown to be generally consistent with a conventional model that uses the log-ratios in a Cox proportional hazards model, and potentially important genes selected by the proposed model only are found to have known connections with breast cancer. That is, conventional models that do not account for measurement error may fail to detect these genes' associations between event and gene expression. In addition to detecting associations, the conventional models may underestimate the strength of these associations because models not accounting for measurement error are known to be biased towards the null, and this bias may be avoided in the proposed model. The model was shown to be robust to some parametric assumptions for inference about the parameter of interest, and the new GEI's are found to be highly correlated with the log-ratios. Further, the model is demonstrated to have good operating characteristics concerning type I and type II error rates as well as accurate coverage of the parameter values by the HPDs. However, the issue of False Discovery Rates (FDR) is not addressed

here. Conceivably, permutation of the survival times could be applied to the data in order to estimate the false discovery rate. Permutation is regularly applied in the case of the Cox model and in other frequentist approaches in microarray data Sorlie et al. (2001), yet such permutations would be not computationally feasible for a Bayesian analysis using this model, and permutation is only valid under exchangeability which excludes more complex models with clinical covariates. The problem of estimating FDR for Bayesian models is one of current research (Efron et al., 2001; Ibrahim et al., 2002; Newton et al., 2004a; Tadesse et al., 2005), and the estimation of the FDR can be obtained by using the mean posterior probability. If one is interested in which genes are most likely to be associated with the time-to-event data, an ordering of the genes in terms of association is required. In the frequentist setting, the p-values for the test statistics can generate the ordering. One may easily derive such an ordering from the model presented here by calculating the posterior probability that $\gamma_g = 0$ as in Tadesse et al. (2005). Overall, this model has an important advantage over the conventional one in that it accounts for measurement error which is a significant additional source of variation.

3 Microarrays and Genetics

The second paper derives an enhanced method for finding associations between genotype and gene expression. Microarrays represent high-dimension complex traits that can be influenced by the genotype of the cells. The purpose of genetic analysis of microarray data is to understand the influence of genotype on gene expression as an intermediary between genotype and the directly observable complex traits such as blood pressure, cholesterol, obesity and disease states like diabetes. Linking genotype and expression may help to elucidate genetic networks as well. Jansen and Nap (2001) asserted that the combined analysis of gene expression and genetic variation be called "genetical genomics". Others have called it eQTL analysis for Expression Trait Loci. eQTL analysis methods are closely related to Quantitative Trait Loci (QTL) methods that have been developed for single or a few traits (Lynch and Walsh, 1998). The genetic analysis of quantitative traits has a very long history dating back to Francis Galton in 1869 (Galton, 1892).

3.1 Fundamentals of Genetics

The basic aim of eQTL and QTL analysis is to find associations between the genotype which is a set of positively correlated, categorical variables and the phenotype that is a continuous response. For a review of QTL methods, see Lynch and Walsh (1998).

Experimental or observational design plays a pivotal role in the analysis techniques used in mapping or detecting eQTL and QTL. The main consideration is whether the

population tested is *inbred* or *outbred*. Inbred populations are those whose parents are closely related. Specifically, *recombinant inbred lines* (RILs or RI strains) are the results of multiple generations of brother-sister mating (Lynch and Walsh, 1998). Through recombination, the offspring will become almost completely homozygous, meaning that the maternal and paternal chromosomes have the same genotypes. The offspring will have identical genotypes except for the differences between sexes. Two RILs can be crossed in different ways depending on the experimental design. For example, F_1 designs compare offspring from the cross of 2 RILs. F_2 designs involve the offspring of the F_1 generation and so on. The *backcross* design compares the cross of the F_1 line with one of the parents. The observational designs of outbred populations are very different from those of inbred populations. Outbred parents and offspring are those whose ancestors are not closely related. This poses additional analytical challenges compared to inbred populations, but many important studies of humans involve outbred subjects. Lynch and Walsh (1998) stress that the outbred designs examine within population trait variability while the inbred designs examine between population variances, and they give the major differences between the two. The variability of genetic markers is not well controlled in outbred populations. For example, markers may not be *informative*, meaning that the genotypes are polymorphic (having variation) for the subjects in the study. On the other hand, outbred parents could have excess variability at a locus. For example, if there are 4 or more genotypes, then the analysis can become less powerful to detect QTLs.

3.2 Fundamentals of QTL Analysis

The analysis of inbred strains involves the comparisons of means of populations. One fundamental idea in QTL analysis is that genotype influences the mean value of the trait y so that

$$y_i = \mu + \beta x_i + \epsilon_i \tag{40}$$

where β is the QTL effect, and x_i is the QTL genotype of the i^{th} subject, and ϵ_i is a random error with variance σ^2. The vectors x_i and β may be two or three dimensional depending on the number of different genotypes and whether or not the effect of the QTL is *additive* or has a *dominance* component. If there are three genotypes say QQ, Qq and qq, then an additive model would have means $\mu_{QQ} + a = \mu_{Qq} = \mu_{qq} - a$ for some a. For dominance models, there is no such a, but there are a and d where $\mu_{QQ} + a = \mu_{Qq} + d = \mu_{qq} - a$. The location of the QTL or eQTL is generally unknown in advance so that markers must be used as proxies. One may use the markers themselves, but for sparse markers, this may have disadvantages such as underestimating the QTL effect and inaccurate estimation of the QTL location. Interval mapping was developed by Lander and Botstein (1989) in order to analyze the possible occurrence of a QTL between the markers. The likelihood for the QTL with additive effect a becomes

$$L(\mu, a, \sigma^2) = \prod_{i=1}^{n} [G_i(0)L_i(0) + G_i(1)L_i(1)] \tag{41}$$

where $G_i(x)$ is the probability of the genotype $x \in \{0, 1\}$ of the i^{th} subject. $L_i(x)$ is the likelihood function given genotype x. The probability $G_i(x)$ may be calculated conditionally upon the distance from the left and right flanking markers for any position between them. The above likelihood can be maximized by the Expectation Maximization (EM)

algorithm for positions uniformly distributed across the genome. This form of interval mapping is computationally intensive, and eQTL would greatly increase those demands. The results of interval mapping can be approximated by another more computationally efficient method of Haley and Knot (1992). They proposed that instead of Equation (40) one substitutes the expected value of x conditional upon the flanking markers for x. This allows one to calculate the likelihood for positions between markers like the interval mapping of Lander and Botstein, but it avoids the burden of the iterative EM algorithm. Another significant advance in QTL involves correction for multiple QTL affecting a single trait. If there are many QTL then the model becomes $y_i = \mu + \sum_g \beta_g x_{gi} + \epsilon_i$ where the subscript g indexes the different QTL. The existence of multiple, linked QTLs is known to bias the effect and location of methods that detect the largest single QTL (Zeng, 1993). Zeng (1993) proposed Composite Interval Mapping (CIM) to overcome bias due to the *composite* effects of multiple QTL. CIM models the trait y as a function of the genotype anywhere in the interval $(j, j+1)$ between the two flanking markers j and $j+1$ and the genotypes at all other markers. The phenotype model becomes $y_i = b_0 + b^* x_i^* + \sum_{k \neq j, j+1} b_k x_{ki} + \epsilon_i$ where k indexes the nonflanking markers. The b^* parameter measures the effect of the loci of interest while the b_k parameters are the effects of the background trait variability due to the k^{th} marker. The number of background markers to adjust for is not given by the model, but the markers $j-1$ and $j+2$ should always be included because all of the other markers on the same chromosome are conditionally independent of x_i^*.

eQTL detection in human studies must use the analytic methods of outbred analysis. The underlying model for the means is the same as for inbred populations, but the

modeling focuses on the analysis of variance components. In the notation of Almasy and Blangero (1998), we have $y = \mu + X'\beta + \sum_{i=1}^{n} \gamma_i + g + \epsilon_i$ where y is the vector of trait values, γ is the effect of QTL i, and ϵ is the random error. The term $X'\beta$ represents the covariates (e.g. age, sex) and the corresponding regression coefficients. The g parameter represents the effects of the *polygene* which is the composite of many QTLs with small effects. The variance of y (σ_y^2) can be written in terms of the independent genetic components $\sum_{i=1}^{n} \sigma_{\gamma_i}^2$ and the random error σ_ϵ^2 so that $\sigma_y^2 = \sum_{i=1}^{n} \sigma_{\gamma_i}^2 + \sigma_g^2 + \sigma_\epsilon^2$. The covariance of any two related individuals is a function of k_{ji} which is the probability that the i^{th} QTL has j alleles that are Identical by Descent (IBD). Ignoring dominance effects, the covariance of these two relatives is $\mathrm{Cov}(y_1, y_2) = \sum_{i=1}^{n} (k_{1i} + k_{2i})\sigma_{ai}^2$. One may make an approximation that $\sigma_a^2 = \sum_{i=1}^{n} \sigma_{ai}^2$, and let $\phi = \frac{1}{2}E[(k_{1i} + k_{2i})]$ where ϕ is called the expected kinship coefficient. This gives $\mathrm{Cov}(y_1, y_2) \approx 2\phi\sigma_a^2$. If one is interested in QTL i only, we have $\mathrm{Cov}(y_1, y_2) = \pi_i \sigma_{ai}^2 + 2\phi\sigma_g^2$ where $\pi_i = k_{1i} + k_{2i}$. The background or *polygenic* effects are reflected by σ_g^2. The term π_i is the probability of an allele of the i QTL being IBD and is called the coefficient of relationship as in Almasy and Blangero (1998). The phenotypic covariance for a general pedigree may be written as $\Omega = \sum_{i=1}^{n} \hat{\Pi}_i \sigma_{ai}^2 + 2\Phi\sigma_g^2 + I\sigma_e^2$ where the matrix $\hat{\Pi}_i$ has elements (π_{ijl}) that indicate the proportion of IBD alleles of QTL i shared by individuals j and l. Φ is the matrix of kinship coefficients. The estimation of the $\hat{\Pi}_i$ matrix is not straightforward for general pedigrees, and there are methods designed for estimating the IBD probability for genetic marker locations (Amos et al., 1990; Davis et al., 1996) and locations in between markers (Fulker et al., 1995; Almasy and Blangero, 1998). Almasy and Blangero developed software for general pedigrees that estimates the Π_i matrix for positions between markers with a regression based approach.

The likelihood for the phenotypes given the form of the covariance matrix is then given by

$$\log(L(\mu, \sigma^2_{ai}, \sigma^2_g, \beta|y, X)) = -\frac{t}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}\Delta'\Omega^{-1}\Delta \tag{42}$$

where $\Delta = y - \mu - X'\beta$. The hypothesis concerning whether or not QTL i exists is equivalent to testing $\sigma^2_{ai} = 0$. Since the test involves the boundary of the parameter space ($\sigma^2_{ai} \geq 0$), the distribution of the likelihood ratio statistic does not have a chi-square distribution. Self and Liang (1987) showed that the likelihood ratio statistic will follow a mixture of chi-square distributions, and this null distribution is the basis for inference.

3.3 eQTL analysis

Most of the existing analyses of and methods for eQTL detection are adaptations of QTL methods applied to gene expression data. The main difference is that the number of traits is much increased. Attempts are made to reduce the number of traits or dimensions that are considered. One motivation is that transcripts with low variation in expression *between* genotypes are not likely to be controlled by eQTLs. The removal of these low variance transcripts is then thought to reduce the number of false positive eQTL detections as in Schadt et al. (2003). Schadt et al. (2003) excluded transcripts of low variability, and then used interval mapping with a likelihood ratio threshold to identify eQTLs for a specific trait. Another motivation for reducing the number of transcripts considered is that some of the expression measurements may not be reproducible *within* an genotype. Carlborg et al. (2005) borrowed the concept of repeatability from

Falconer and Mackay (1996). Repeatability is defined as the ratio of within line variance to between line variance. Carlborg et al. (2005) showed that when transcripts with low repeatability are excluded, the power for detection of eQTLs is increased. However, as Schadt et al. (2003) and Chesler et al. (2005) point out, excluding subsets of transcripts may increase the number of false negatives or non-discoveries. Instead of excluding sets of transcripts, the dimension of the phenotype may be reduced by combining transcripts to form a smaller number of "supertraits." Yvert et al. (2003) reduced the number of individual traits by clustering groups of genes that were significantly correlated, and the mean expression level of transcripts in the cluster was the quantitative trait considered. Lan et al. (2003) proposed that the principle components of the gene expression data and clusters based on transcripts of interest could be used as supertraits to improve the power of eQTL detection.

The attempts to control false discovery or Type I error rates in eQTL analyses have mostly been derivative of the methods applied to QTL framework. Schadt et al. (2003), Morley et al. (2004), and Monks et al. (2004) used a genome-wide p-value cutoff based on a Bonferroni correction to all possible eQTL associations. Chesler et al. (2005), Hubner et al. (2005), and Carlborg et al. (2005) used a combination of a genome-wide permutation p-value and the FDR estimation method of Storey (2002). The permutation procedure of Churchill and Doerge (1994) was applied to each transcript. This procedure simply permutes the trait (transcript) values, and calculates the test statistic for each loci on the genome. The maximum test statistic across the genome for each permutation gives an empirical null distribution. The observed test statistic is then given a p-value according to this null distribution. In these eQTL experiments, this p-value is then entered into

Storey's q-value procedure so that the FDR can be computed.

Very recently, Storey et al. (2005) have developed a method that uses a forward selection process to create a multiple eQTL model, meaning that many loci affect the mean expression level of a transcript. The method proceeds as follows. First, for each transcript, the loci on the genome with the highest likelihood of association is chosen using a method related to Efron (2004) that calculates an empirical Bayes estimate of posterior probability for association. Next, the effect of the chosen loci is included in the model for each corresponding transcript, and the procedure selects the loci with the highest probability of association with the transcript given the loci selected in the previous stage. The posterior probability that both loci selected in the two rounds of the procedure are associated with the transcript is given by

$$\begin{aligned} P\{\text{loci 1 and 2 are associated}|\text{Data}\} &= P\{\text{locus 2 associated}|\text{Data, locus 1 associated}\} \\ &\quad \times P\{\text{locus 1 associated}|\text{Data}\}. \end{aligned} \tag{43}$$

This methods allows for the FDR to be calculated as the average posterior probability of no association ($1 - P\{\text{loci 1 and 2 are associated}|\text{Data}\}$) for the selected subset of traits with two loci models. The forward selection can be applied for more than two levels, but forward selection procedures are generally known to be biased. Whether or not this bias affects the biological inferences is not known at this point.

3.4 Data Structure

The motivating dataset that we consider is from an experiment on brain tissue from mouse recombinant inbred (RI) strains with Affymetrix microarray measurements. The

microarray dataset is available from www.genenetwork.org by searching for mouse data, BXD group, and whole brain tissue. RI strains are the product of multiple generations of inbreeding that result in offspring that are homozygous for either of the two founding parents. BXD (B strain crossed with D strain) strains have homozygous alleles of either one of two parents B (C57BL/6J) and D (DBA/2J). The BXD panel and RI panels in general have several advantages for eQTL mapping experiments (Chesler et al., 2004). The RI mouse model has been broadly used for the genetic exploration of complex diseases and as a model of human disease (Chesler et al., 2004). The RI strains are a renewable source of genetically identical animals that can be used in experiments that are reproducible from laboratory to laboratory. The genetic identity reduces the need for genotyping and facilitates exploring gene by environment interactions. Also, there are continuously evolving databases for comparison and integration of results. In this paper, the model is applied to RI designs, but it can be applied to many breeding designs such as backcross or F2. The basic goal of eQTL detection is to find associations between the categorical predictors (genotypes) and the continuous response (transcript expression level). (For a general reference on QTL analysis, see Lynch and Walsh (1998)). In the most simple circumstance and in the case of RI strains, the genotypes are of two varieties denoted as 0 or 1. The genotype of any individual k out of n is then a vector of length M of 0's and 1's where M is the number of genetic markers. The index for the transcript will be t out of T. The microarray measurement for a transcript t is given by $\boldsymbol{y}_t$ so that

we have

$$\begin{aligned} \boldsymbol{y}_t &\equiv \text{Vector of gene expression measurements for transcript } t. \\ \boldsymbol{y}_{t,m,g} &\equiv \text{Subvector of } \boldsymbol{y}_t \text{ for individuals having genotype } g \text{ at marker } m. \end{aligned}$$

The transcripts can be either equivalently expressed (EE) meaning that genotypes are not associated with expression level or differentially expressed (DE) meaning that one or more marker loci are associated with the transcript's expression.

3.5 The Mixture Over Markers Model

The False Discovery Rate (FDR) is often used to account for the multiple testing problem in microarray analyses, but the multiple testing methods typically applied to microarray data like the q-value method can be anticonservative in the eQTL setting as Kendziorski et al. (2005) point out. The reasons for the shortcoming of these methods are manifold, but the main reason is that FDR methods developed for microarrays typically consider only one alternative hypothesis. That is, the alternative hypothesis is that the transcript is differentially expressed with respect to or correlated with a single biological state. In eQTL analyses, any given transcript has alternative hypotheses for each marker. That is, a transcript could be associated with marker 1, 2, 3, etc. This adds another dimension to the multiple testing problem. In QTL analyses, the existence of merely an association between a marker and a trait is not a sufficient discovery because the QTL discoveries are optimally localized to minimal regions of highest association, not merely some association. Kendziorski's MOM method (Kendziorski et al., 2005) has the advantage that it simultaneously estimates posterior probabilities of all of the MT possible associations be-

tween transcripts and markers. The result of fitting the MOM model gives the posterior probability of a transcript being associated with a certain marker based on the observed data. This has at least three advantages. First, by averaging the posterior probabilities, one can estimate the FDR. Second, for every transcript, one has the conditional probabilities for an eQTL across all of the markers. Returning to the previous example, the MOM model gives the posterior probabilities that transcript t is associated with markers 1, 2, or 3, simultaneously. Third, the model pools information to estimate parameters common to all transcripts.

The MOM model does have some disadvantages. First, the MOM model assumes at most one locus explicitly controls the expression of a given transcript which may not hold for some transcripts. In the QTL literature, there are some methods that are developed which can estimate the number as well as the location of QTLs, but these methods may not be readily applied in the eQTL setting because of the sheer computational burden (Storey et al., 2005). We propose an extension of the MOM method to model two eQTLs conditionally upon finding the first major eQTL. Second, the MOM method assumes that the error variances are equal within predetermined transcript clusters. We relaxed this assumption and suggest a discrete uniform prior for the standard deviations of the errors.

When considering all possible MT associations between transcripts and markers it could be advantageous to utilize the patterns of these associations. Chesler et al. (2004) and Carlborg et al. (2005) noted that transcripts are more likely to be associated with markers that correspond to the genomic location of the transcript. eQTLs that are associated with the transcripts of nearby genes are called *cis* acting eQTLs while eQTLs

that are associated with transcripts that are some distance away are called *trans* acting eQTLs. There are biological reasons for the prevalence of *cis* eQTLs. Namely, the local DNA sequence may contain elements that affect the transcript's regulation or subtly affect the transcript's function (Doss et al., 2005). This biology is one of the major motivations for developing the proposed extension of the MOM method to include genomic locations of the transcripts and markers. There is a possibility that a putative *cis* eQTL is an artifact of sequence variation in the transcript affecting the expression measurement process. However, Doss et al. (2005) examined putative *cis* acting eQTLs in mice and experimentally confirmed that a majority of these associations corresponded to true eQTLs.

First, we describe the MOM model. The marginal likelihood for EE transcripts is

$$f_0(\boldsymbol{y}_t) = \int f_*(\boldsymbol{y}_t|\sigma^2)\pi_{\sigma^2}(\sigma^2)d\sigma^2, \tag{44}$$

where

$$f_*(\boldsymbol{y}_t|\sigma^2) = \int \prod_k f_{obs}(y_{t,k}|\mu_t, \sigma^2)\pi_\mu(\mu_t)d\mu_t, \tag{45}$$

and

$$f_{obs}(y_{t,k}|\mu_t, \sigma^2) = \phi(y_{t,k}; \mu_t, \sigma^2) \text{ and } \pi_\mu(\mu_t) = \phi(\mu_t; \mu_0, \tau_0^2). \tag{46}$$

The transcription measurements are considered independent sampling units. The term $f_{obs}(y_{t,k}|\mu_t, \sigma^2)$ is the distribution of $y_{t,k}$ (the k^{th} element of $\boldsymbol{y}_t$) conditional on σ^2 and the mean μ_t, which for EE transcripts, is the same for all subjects. f_{obs} represents residual, non-genetic variation and genetic variation not explained by the model. The parametric form of f_{obs} is $\phi(y_{t,k}; \mu_t, \sigma^2)$ where $\phi(x; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. One may integrate out the μ_t parameter and find that f_* is a multivariate

normal density with mean μ_0 and a compound symmetric variance $I\sigma^2 + 11'\tau_0^2$ where 1 and I are the vector of ones and the identity matrix respectively.

In Equation (44), the prior for σ^2 is denoted as π_{σ^2}. In Kendziorski et al. (2005), there is no prior on σ^2, but the error variance is chosen to be equal within clusters. These clusters are chosen using the k-means algorithm before MOM is applied. We argue that there is no *a priori* reason to assume that genes have equal variances because they are correlated. Further, the uncertainty in the clustering of genes is not accounted for when the variance categories are fixed before the model fit. We choose a discrete prior for σ^2 so that σ is uniformly distributed over the interval $[0, \sigma_*]$ to cover the range of probable variances where σ_* is determined by estimating the variances of a subset of genes thought to be EE. Equation (44) implies that f_0 is a scale mixture of compound symmetric densities. Differentially expressed (DE) transcripts associated with marker m would have density $f_m(\boldsymbol{y}_t) = \int f_*(\boldsymbol{y}_{t,m,0}|\sigma^2) f_*(\boldsymbol{y}_{t,m,1}|\sigma^2)\pi_{\sigma^2}(\sigma^2)d\sigma^2$ where the corresponding f_{obs} are centered around a different mean ($\mu_{t,m,0}$ or $\mu_{t,m,1}$) according to the genotype of marker m. This implies that f_m is a scale mixture of block diagonal compound symmetric normal densities.

The status of whether or not the transcript is differentially expressed and if so, which marker(s) is it associated with is not known a priori, and this is treated as missing data. In a fully Bayesian context, one would estimate the joint posterior distribution for both the parameters and the missing data. Although such joint estimation would be ideal, the number of missing data points is MT which makes such fully Bayesian estimation computationally infeasible using MCMC methods. As a computationally feasible alternative, we consider an Empirical Bayes procedure in which the parameters are estimated

by maximizing the marginal likelihood. If we consider only marker m and let p_m be the probability of a transcript t being associated with marker m, then the marginal distribution of the data is given by $L_{t,m} = p_m f_m(\boldsymbol{y}_t) + (1 - p_m) f_0(\boldsymbol{y}_t)$. This model may be extended over many markers. We let p_0 be the prior probability of the transcript mapping nowhere (i.e., the null hypothesis that the transcript is not associated with any marker) and equate the mixing proportions p_m with prior probabilities for the marker being an eQTL for a transcript. One may notice that the prior probabilities p_m for a transcript mapping to a particular marker are not dependent on the particular transcript. The likelihood now becomes $L_t = p_0 f_0(\boldsymbol{y}_t) + \sum_{m=1}^{M} p_m f_m(\boldsymbol{y}_t)$ where M is the total number of markers considered. So the likelihood for all transcripts and markers is $L = \prod_{t=1}^{T} L_t$. The model is fitted with the expectation maximization (EM) algorithm. EM is used because the binary (0 or 1) $(M+1) \times T$ dimension matrix Z of random variables $(z_{m,t})$ that determine which marker, if any, is associated with the transcript t are not observed. The case $z_{0,t} = 1$ implies that transcript t is not controlled by any marker, and the case $z_{m,t} = 1$ implies that transcript t is controlled by marker m. Thus, if the $z_{m,t}$ were observed, we would have the complete data. The columns z_t of the matrix of $z_{m,t}$ are multinomial random variables which contain exactly one 1 when observed and have elements whose expectation sum to unity. This model considers the existence of only one major eQTL. The complete data log-likelihood for a given transcript can be rewritten in terms of these $z_{m,t}$ as

$$l_t = \sum_{m=0}^{M} z_{m,t} \log(p_m) + \sum_{m=0}^{M} z_{m,t} \log(f_m(\boldsymbol{y}_t)). \tag{47}$$

The above equation illustrates that the mixture probabilities p_m can be thought of as

the prior component with the $f_m(\boldsymbol{y}_t)$ being the likelihood component. This equation also indicates that z_t has multinomial probabilities $p_0 \ldots p_M$ that can be estimated directly with the EM algorithm by substituting the expectation of $z_{m,t}$. The expectation of $z_{m,t}$ is an important quantity that is equal to the posterior probability of a transcript mapping to the marker m and is given by

$$E[z_{m,t}] = \frac{p_m f_m(\boldsymbol{y}_t)}{\sum_{m=0}^{M} p_m f_m(\boldsymbol{y}_t)}. \tag{48}$$

3.6 Extensions of the MOM model

We describe the proposed extensions of the Mixture Over Marker model, the model fitting procedure, and the calculations of the FDR.

3.6.1 Proximity Model

We extend the MOM model to allow the prior probabilities of a transcript mapping to a marker to depend on the transcript's genomic proximity to the marker. We choose a simple and reasonable relationship to model these prior probabilities. We use a log-linear model for the mixture probabilities $p_{m,t}$ that contains the Kendziorski model as a special case. The responses in the model are the multinomial columns z_t of the Z matrix that can be converted into a $(M+1) \times (M+1)$ contingency table ζ with elements $\zeta_{ij} = \sum_{t \in \mathrm{C}_i} z_{j,t}$ where $\mathrm{C}_i \equiv \{t \quad | \quad t \text{ closest to marker } i\}$. The element ζ_{ij} is the number of transcripts closest to marker i that map to marker j where $i, j \in \{0, 1, \ldots, M\}$. The 1^{st} row of ζ (ζ_{0j}) represents transcripts that were not sufficiently close to any marker. The 1^{st} column of the ζ (ζ_{i0}) represents the transcripts mapping to no marker. Converting Z into ζ allows

the reduction of the data from $(M+1) \times T$ elements of Z to the $(M+1) \times (M+1)$ elements of ζ because T is often several times larger than M. A log-likelihood is derived under the assumption that the ζ_{ij}'s are a vector of random variables that have a Poisson distribution. This generalized linear model with the canonical link is proportional and equivalent to a multinomial likelihood for the z_t's. The linear, systematic component of the model for the elements of the table is

$$\log(E[\zeta_{ij}]) = \nu_{ij} = \alpha_i + \beta_j + \gamma I[i = j] I[i > 0] \tag{49}$$

where $I[\ \ ]$ is the indicator function. The $I[i > 0]$ term exists because the first row of the table corresponds to transcripts that are not close to any marker. The parameters $\alpha_0 \ldots \alpha_M$ are nuisance parameters that model the row totals so that the log-linear model and the multinomial models are equivalent. The parameters $\beta_0 \ldots \beta_M$ correspond to the marker specific effects. The β_0 parameter is related to the log prior probability that a transcript is not associated with any marker. Because the multinomial probabilities must sum to 1, one of the β_m parameters is determined by the others, and without loss of generality, we set β_0 to 0. The size of β_m varies greatly. Some markers do not appear to regulate any transcripts while other markers might modulate hundreds of transcripts. The γ parameter represents the effects of proximity and is the increase (if $\gamma > 0$) in prior probability of a transcript being associated with the closest locus. Explicitly, the prior probability for $z_{m,t}$ is

$$p_{m,t}(\beta, \gamma) = \frac{\exp(\beta_m + \gamma I[m \text{ is closest marker to } t])}{\sum_{m'=0}^{M} \exp(\beta_{m'} + \gamma I[m' \text{ is closest marker to } t])}. \tag{50}$$

It is clear that if we fix t then $\sum_{m=0}^{M} p_{m,t} = 1$. If $\gamma = 0$, then the log-linear model becomes equivalent to the Kendziorski's multinomial model for the mixture proportions

$z_{m,t}$. The modeling of the mixture proportions adds some numerical difficulties because some markers (say marker m) may not be associated with any transcripts which precipitates convergence problems because this implies $\beta_m \to -\infty$. This may be alleviated by using a normal prior on the β parameters. These priors may be easily implemented by the method of Knuiman and Speed (1988) who developed these models to utilize prior information for contingency tables, but they noted that these priors also accommodate fitting with low frequency cells in tables. We chose a diffuse prior such that $\beta \sim N(0, \sigma_\beta^2)$ which is both a sensible prior and is equivalent to using a small ridge parameter that we denote as $\lambda = \frac{1}{2\sigma_\beta^2}$. These normal priors can be fit by adding a penalty parameter to the iterative weighted least squares algorithm. We observed that the inferences were insensitive to the prior choice of λ. The model performed well for values of λ between 10^{-2} and 10^{-4} which imply a large prior variance for β_m.

It is worth mentioning that the proximity model is identifiable when the MOM model is not. For example, if there are two or more flanking markers with identical genotypes, then the MOM model is incapable of distinguishing between them so that the prior probabilities of these markers would be nonidentifiable. In the proximity model, the markers would be differentiated by their proximity to different transcripts, and the posterior probability for being an eQTL would be highest for the closest markers.

3.6.2 Model Fitting

We use a variant of the EM algorithm (Dempster et al., 1977) known as the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). The full log

likelihood conditional on $z_{m,t}$ is

$$l = \sum_{t=1}^{T} \sum_{m=0}^{M} z_{m,t} \log(p_{m,t}(\beta, \gamma)) + \sum_{t=1}^{T} \sum_{m=0}^{M} z_{m,t} \log(f_m(\boldsymbol{y}_t|\boldsymbol{\Omega})). \tag{51}$$

A convenient aspect of this model is that the β and γ parameters can be estimated independently of $\boldsymbol{\Omega} = (\mu_0, \tau_0^2)$ conditionally on the expectation of $z_{m,t}$. The expectation of $z_{m,t}$ is

$$E[z_{m,t}] = \frac{p_{m,t}(\beta, \gamma) f_m(\boldsymbol{y}_t)}{\sum_{m=0}^{M} p_{m,t}(\beta, \gamma) f_m(\boldsymbol{y}_t)}. \tag{52}$$

The ECM algorithm begins by choosing initial values for all of the parameters. The expectation of $z_{m,t}$ is then calculated. Next, we maximize the expected log-likelihood over the β and γ parameters. There are a few considerations when fitting this model that were not previously mentioned. First, the transcripts should be ordered by their genomic locations to facilitate the collapsing of Z into the table of ζ_{ij}. Second, sparse matrix operations effectively reduce the computation time. The dimension of the design matrix is $\sim M^2 \times 2M$ which might cause the usual generalized linear model fitting procedures to fail because of computer memory and time limitations. We used the sparse matrix operations package developed by Koenker and Ng (2003) to implement the iteratively weighted least squares algorithm with the λ ridge parameter to fit the generalized linear model. The next step of the algorithm is recomputation of the expected value of $z_{m,t}$. Then, the $\boldsymbol{\Omega}$ parameters are maximized. We maximize these parameters with a generic optimization algorithm provided by the R nlm function (R Development Core Team, 2004b). The ECM algorithm continues alternately conditioning on $\boldsymbol{\Omega}$ and the β and γ parameters until convergence.

3.6.3 Calculation of the False Discovery Rate

We estimate the FDR in a similar manner as Kendziorski et al. (2005) and Newton et al. (2004b). The determination of the FDR depends on the calculation of the posterior probabilities of EE ($p_{EE,t}$) and DE ($p_{DE,t} = 1 - p_{EE,t}$) for each transcript. This is given by

$$p_{EE,t} = E[z_{0,t}] = \frac{p_{0,t} f_0(\boldsymbol{y}_t)}{\sum_{m=0}^{M} p_{m,t}(\beta, \gamma) f_m(\boldsymbol{y}_t)}. \tag{53}$$

To control the FDR to be less than α, we choose a threshold $\kappa(\alpha)$ for the $p_{EE,t}$ so that if $p_{EE,t} < \kappa(\alpha)$ then the transcript is identified as DE or mapping to some marker. For any given $\kappa(\alpha)$, the FDR is

$$\text{FDR} = \frac{\sum_{t=1}^{T} p_{EE,t} I[p_{EE,t} < \kappa(\alpha)]}{\sum_{t=1}^{T} I[p_{EE,t} < \kappa(\alpha)]} \leq \alpha. \tag{54}$$

The FDR is the average posterior probability of being EE of those transcripts that are selected as DE. In most nontrivial cases, some of the $p_{EE,t}$ that are averaged will exceed the FDR. We have seen in real data analyses that controlling FDR alone could lead to a poor decision rule which results in a $\kappa > \frac{1}{2}$. This implies that some transcripts declared to be DE may have $p_{EE,t} > \frac{1}{2}$. Thus, we advocate controlling κ, but the FDR remains a useful summary measure of the overall reliability of a set of inferences. Particularly, the FDR is useful for comparing the average reliability of lists generated by different methods as is done with the simulations of Section 3.7.

3.6.4 Multiple eQTL extension of the MOM model.

Multiple eQTL may be discovered using the following extension based on a forward model selection process. We consider the 2 eQTL case. In the first stage, one applies the MOM

or the Proximity Model to select the most likely associated marker for each transcript identified as DE. Call these markers m_t^*. In the second stage, one fits a variant of the MOM model for the case of 2 eQTLs which we describe below. We condition on the event that the transcript t has at least two means based upon the genotypes of marker m_t^*. Let $\boldsymbol{y}_{t,m_i,g_i,m_j,g_j}$ be the subvector of $\boldsymbol{y}_t$ that contains observations from individuals having the genotypes g_i and g_j for markers m_i and m_j respectively. The density for a transcript mapping to 2 markers (m_t^* and another marker m) is given by

$$f_{m_t^*,m}(\boldsymbol{y}_t) = \int f_*(\boldsymbol{y}_{t,m_t^*,0,m,0}|\sigma^2) f_*(\boldsymbol{y}_{t,m_t^*,0,m,1}|\sigma^2) f_*(\boldsymbol{y}_{t,m_t^*,1,m,0}|\sigma^2) f_*(\boldsymbol{y}_{t,m_t^*,1,m,1}|\sigma^2) \pi_{\sigma^2}(\sigma^2) d\sigma^2. \tag{55}$$

This is a natural extension of the definition of $f_m(\boldsymbol{y}_t)$ in Section 3.5 and represents a 2 eQTL model with 4 different means corresponding to the 4 different genotype combinations. The marginal likelihood of the transcript data becomes $L_t = p_1 f_{m_t^*}(\boldsymbol{y}_t) + \sum_{m \neq m_t^*} p_2 f_{m_t^*,m}(\boldsymbol{y}_t)$ where p_1 is the prior probability of being associated with one marker only, and $p_2 = (1-p_1)/(M-1)$ is the prior probability of mapping to any one of the additional $M-1$ markers. One may fit this model using the EM algorithm as described. The posterior probability of a transcript being only associated with marker m_t^* is denoted as $p_{EE_2,t} \equiv E_2[z_{m_t^*,t}]$, and can be estimated by

$$E_2[z_{m_t^*,t}] = \frac{p_1 f_{m_t^*,m_t^*}(\boldsymbol{y}_t)}{p_1 f_{m_t^*,m_t^*}(\boldsymbol{y}_t) + \sum_{m \neq m_t^*} p_2 f_{m_t^*,m}(\boldsymbol{y}_t)}. \tag{56}$$

Controlling the FDR in the setting of forward model selection is possible by calculating the conditional probabilities using the method of Storey et al. (2005). The FDR of the second stage is calculated based upon the posterior probabilities of the second stage

conditionally upon the posterior probabilities in the first stage. That is,

$$\begin{aligned} & P\{m_t^* \text{ and additional marker are associated with } t|\text{Data}\} \\ = & P\{\text{Additional marker associated}|\text{Data, marker } m_t^* \text{ associated}\} \\ & \times P\{\text{marker } m_t^* \text{ associated}|\text{Data}\} \\ = & (1 - p_{EE_2,t}) \times E[z_{m_t^*,t}]. \end{aligned} \tag{57}$$

where $E[z_{m_t^*,t}]$ is calculated as in Equation (48). The FDR of the second stage is the average of $1 - P\{m_t^*$ and additional marker are associated|Data$\}$ for the selected subset of transcripts with two loci models. Fitting more than 2 eQTLs proceeds with similar arguments.

3.7 Simulated Data Analysis

We performed analyses under the controlled setting of simulated data in order to explore the operating characteristics of the proposed proximity model in comparison to the previous MOM model. The comparison is between the realized performance of both methods in terms of power and FDR control. The definition of power and false discovery rate for eQTL analyses is not trivial and should be defined. We define power to be the probability of identifying a transcript as DE with the posterior distribution of an eQTL having a maximum at the true location of an eQTL. That is, power is the probability of detecting the true eQTL and localizing it to the correct position. We define the realized FDR to be the probability of the union of two mutually exclusive events. First, the model declares a transcript is associated with any marker when the transcript is independent of all markers. Second, for those transcripts associated with a marker, the model declares

the transcript to be associated with some marker, but neither the true eQTL nor the flanking markers are in the 90% Highest Posterior Density (HPD) region. We defined this HPD region as the minimal set of markers whose combined posterior probabilities of association with the transcript is greater than or equal to 90%. The formula for calculating the realized FDR is given by

$$\frac{\#\text{ of Transcripts Falsely Identified as DE} + \#\text{ of Transcripts Incorrectly Localized}}{\#\text{ of Transcripts Identified as DE}}.$$

The simulation experiment details follow. A total of 500 datasets were simulated with 100 datasets for each value of $\gamma = [0.0, 2.0, 4.0, 6.0, 8.0]$. These values of γ correspond to increases in prior probability of a transcript being controlled by the closest marker. The corresponding proportions of DE transcripts that were controlled by the closest marker were $[3\%, 17\%, 59\%, 91\%, 98\%]$. Each dataset had 500 transcripts and 30 markers equally distributed amongst 3 chromosomes evenly spaced with a 10 cM distance between. The transcripts had a 0.6 probability of being EE, and the eQTL were uniformly distributed across markers. The genotypes of $n = 60$ individuals were simulated once and held fixed for all datasets. Realistic distributions were chosen for π_μ, π_{σ^2}, and f_{obs}. To generate these distributions, the MOM model was applied to the BXD dataset described in the next section. We used this model to find a subset of about $5,000$ transcripts that were EE. For π_μ, we resampled the distributions of the sample means of these transcripts. For π_{σ^2}, we independently resampled the distribution of the corresponding sample variances, and the non-genetic error distribution f_{obs} was simulated by independently resampling the empirical deviations of the EE transcripts about their sample means. To fit our model, we chose σ_* to correspond to the 99% quantile of π_{σ^2} and chose 10 points uniformly within

this range for the discrete prior. We fixed our false discovery rate at 0.05 as calculated by Equation (54). The ridge parameter was selected to be $\lambda = 0.05$ which implies that $\beta \sim N(0, 10)$ and gave stable results which converged quickly. For comparisons against a naive method, we used a t-test comparing the means of the genotypes at each marker. We then calculated the q-values of all MT tests. We declared a relationship between marker and transcript if the q-value was < 0.05. The power of the t-test approach is defined to be the probability of finding the true eQTL at the marker having the minimal q-value. The realized FDR is similarly defined as the probability of the union of two events for a transcript. The first event is declaring a transcript to be associated with some marker when it is independent of all markers. Second, for transcripts that have an eQTL, the transcript is declared to be DE, but the true eQTL and flanking markers all have q-values that are > 0.05.

The results are shown in Figure 6 and Figure 7. In Figure 6, there are three power plots representing the overall power to detect eQTLs, the power to detect the *cis* subset of eQTLs, and the power to detect the *trans* subset of eQTLs.

The similarities in power occur because the models are closely related, however, one may see that the overall power of the proximity model is higher than the power of the MOM model as γ becomes large. It may be surprising to see that the power of the proximity model is maintained even if the γ parameter is 0. This is because the γ parameter is estimated to be close to zero when $\gamma = 0$ in the proximity model. As one might expect, the power advantage of the proximity model is increasing for increasing values of γ, and the differences in posterior inferences are substantial for a subset of transcripts as will be seen in the next section. The power advantage appears to be due to

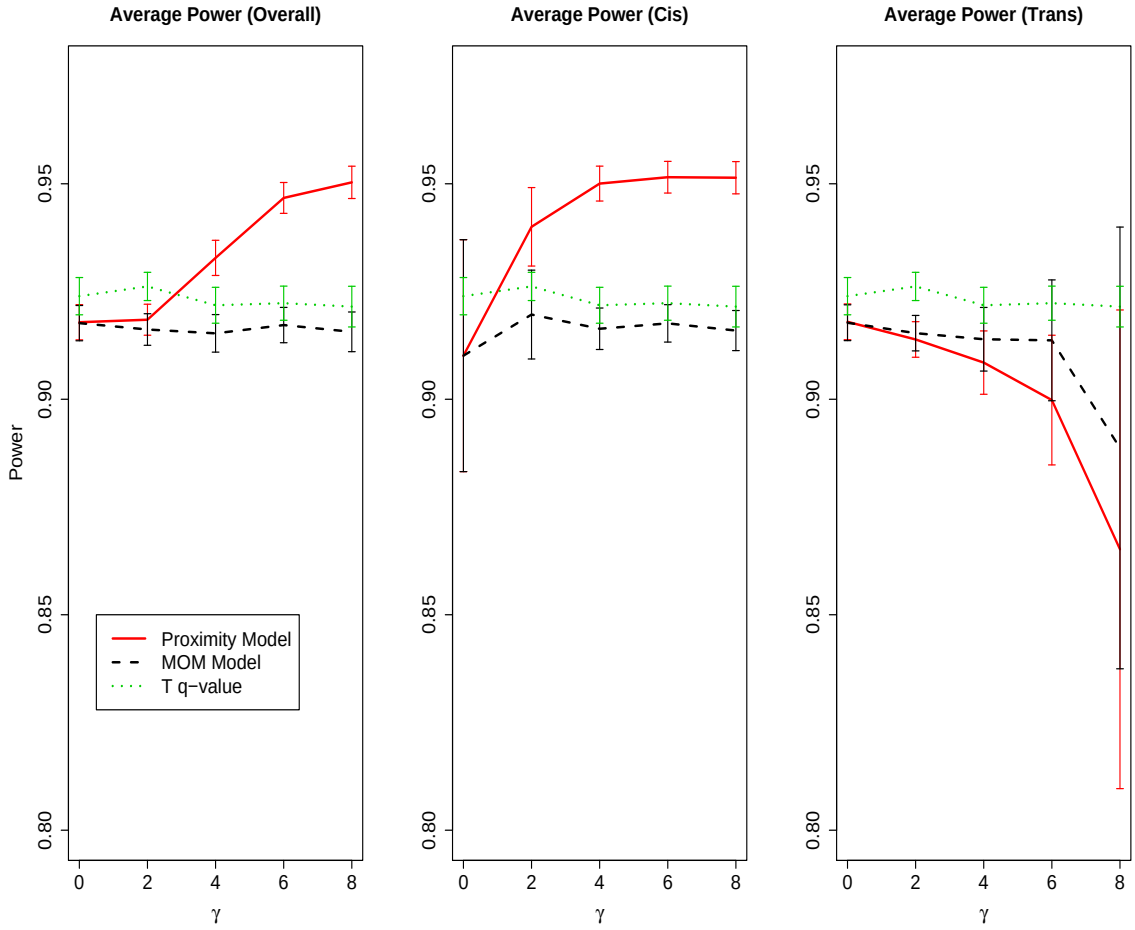

Figure 6: Simulated Data: Power Comparisons with 95% confidence intervals as a function of the proximity effect γ. The overall power is shown on the left. The power plots by *cis* transcripts and *trans* transcripts are middle and right.

the prevalence of *cis* transcripts, but the MOM model has a more variable corresponding power advantage for detecting *trans* eQTLs. The FDR comparison is shown in Figure 7.

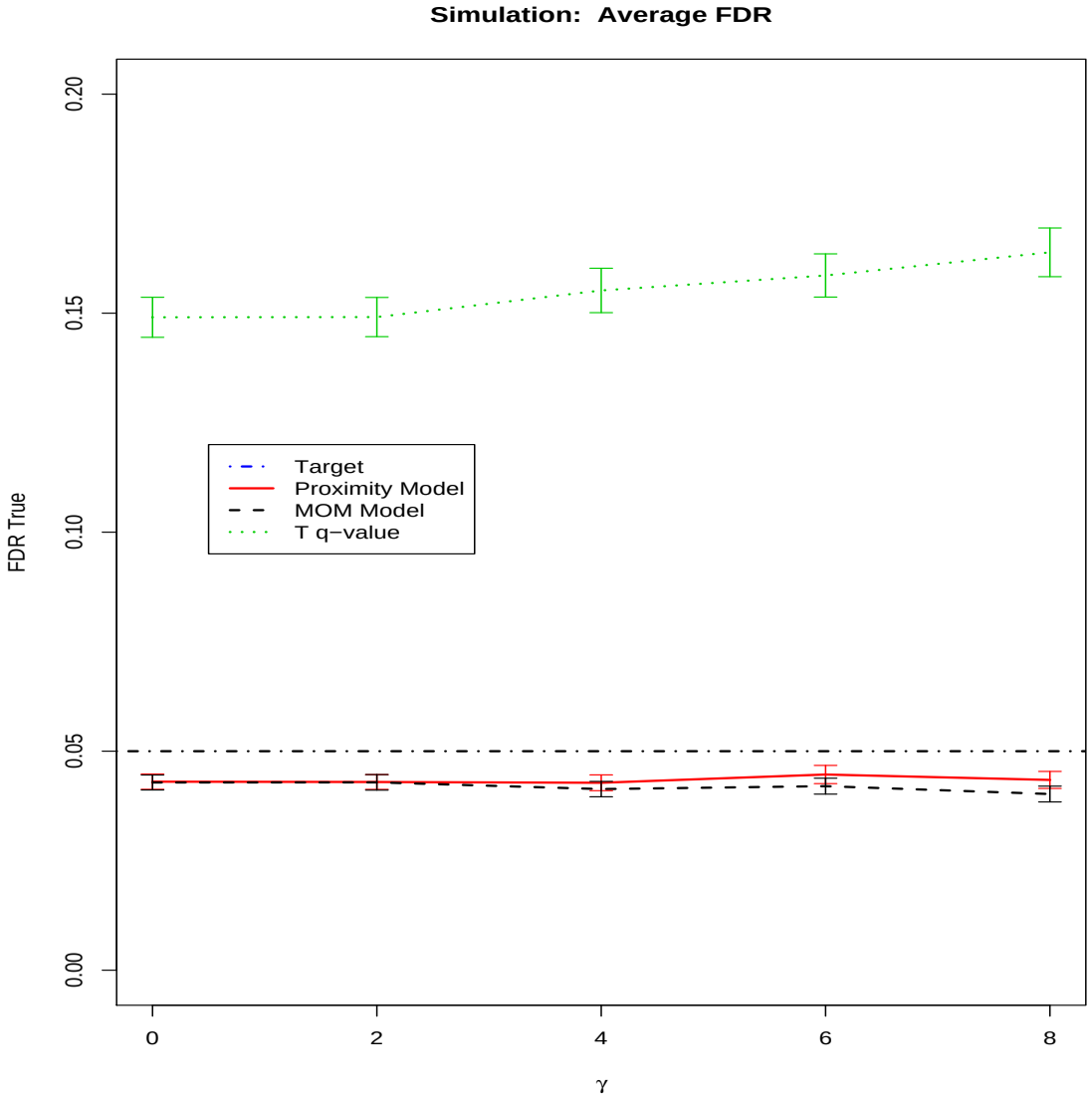

Figure 7: Simulated Data: FDR Comparisons with 95% confidence intervals as a function of the proximity effect γ.

The naive method called "T q-value" has similar power to the MOM method, but it controls the FDR very poorly with a mean FDR of 0.155 and 95% CI $(0.114, 0.197)$ which is much higher than the target FDR of 0.05. This result is consistent with

Table 7: Simulated Data Parameter Estimates

	Parameter	Sample	$\gamma = 0$	$\gamma = 2$	$\gamma = 4$	$\gamma = 6$
MOM	μ	5.848	5.85 (0.06)	5.85 (0.06)	5.85 (0.06)	5.85 (0.06)
Model	τ_0	1.615	1.61 (0.04)	1.61 (0.04)	1.62 (0.05)	1.62 (0.04)
Proximity	μ	5.848	5.85 (0.07)	5.85 (0.06)	5.85 (0.06)	5.85 (0.06)
Model	τ_0	1.615	1.61 (0.04)	1.61 (0.04)	1.62 (0.05)	1.62 (0.04)
	γ	-	-0.14 (0.60)	1.99 (0.21)	4.05 (0.16)	5.99 (0.30)

Table 8: Simulated Data Parameter Estimates Continued

	Parameter	Sample	$\gamma = 8$
MOM	μ	5.848	5.86 (0.06)
Model	τ_0	1.615	1.62 (0.05)
Proximity	μ	5.848	5.86 (0.06)
Model	τ_0	1.615	1.62 (0.05)
	γ	-	7.42 (0.33)

Kendziorski's observations. The average realized FDR of the proximity model is 0.0434 with 95% CI $(0.0303, 0.0623)$ compared to the MOM average FDR of 0.0419 with 95% CI $(0.0235, 0.0587)$. Tables 7 and 8 show the estimates of the remaining parameters with their standard deviation over the simulated datasets.

The μ_0 and τ_0 parameters were estimated accurately by both models. The γ parameter was only estimated by the proximity model, and one may notice accurate estimation.

The 2 eQTL model was fit to 100 datasets simulated with the same distributional

assumptions and genotypes as above. There was no effect of proximity ($\gamma = 0$), and $n = 60$. The probability of EE was 0.6, and the probabilities of a transcripts being associated with one or two eQTLs were 0.2 and 0.2, respectively. The positions of the eQTLs were chosen independently of one another with uniform probability throughout the genome. The two stage fitting procedure was applied, and the threshold κ for $p_{EE,2}$ and p_{EE} was set to 0.05. This method identified both eQTLs at the modes of the posterior distributions 72% of the time. The FDR for declaring associations with these 2 eQTLs was defined as having the posterior mode of the 2 eQTLs outside of the flanking markers for both eQTLs or declaring associations when there are none. The average FDR was 0.016 with a 95% CI of [0.0,0.046], and the average nominal FDR was estimated to be 0.049 so that the realized FDR is conservatively controlled.

3.8 Case Study: BXD Dataset

There were 32 BXD strains considered and each of the strains included were sampled from 1 to 4 times for a total of $n = 88$ mice. The Affymetrix U74Av2 chip was used which contained a total of $T = 11,935$ transcripts with known percent identity to genome locations that are available from Affymetrix (www.affymetrix.com). We chose the transcript to be located at the sequence with highest identity. The 277 markers were used with a median spacing of 4 cM. The markers were located on chromosomes 1 to 19 and on the X chromosome. The determination of which marker was closest was made using the Build 33 distances between the transcripts and the markers. Build 33 refers to the version of the mouse genome map used to locate the transcripts and markers. If the tran-

Table 9: BXD Analysis: Comparison of Parameter Estimates from Mouse Experiment Data with the Standard Deviations in ().

Parameter	MOM Model	Proximity Model
$\widehat{\mu_0}$	6.1439 (0.00003)	6.1443 (0.00006)
$\widehat{\tau_0}$	1.5372 (0.00002)	1.5371 (0.00006)
$\widehat{\gamma}$	-	3.6070 (0.008)

script was further than 5×10^6 base pairs away from any marker then it was considered to be not close to any marker. Only 12% of transcripts were not close to any marker by this standard. Some of the markers had missing data which accounted for 1% of all genotypes. We imputed this missing data by sampling genotypes conditionally upon the flanking markers as in Lynch and Walsh (1998). The variability between imputations of the final results was very small and affected $< 0.1\%$ of transcript inferences because of the high density of the markers leading to 99.7% agreement between the imputations of the genotypes. Because of this small variability and computational time constraints, we performed 5 imputations and report the average result.

The raw microarray data was preprocessed with the Robust Multichip Average (RMA) (Irizarry et al., 2003) software to summarize the probe level data and normalize the microarrays, and this output was used as the expression level for each transcript. The MOM method was applied using our implementation. We chose σ_* as discussed in the simulation section and chose 30 points uniformly within this range for the discrete prior.

Next, we applied our extension of the MOM method to the data, and we compared our results. Table 9 shows the estimates for the $\mathbf{\Omega}$ and γ parameters.

Table 10: BXD Analysis: Comparison of Differentially Expressed Genes Parameter Estimates from Mouse Experiment Data

Posterior Probability Threshold $\kappa = 0.05$

		Proximity Model		
		DE	EE	Total
MOM	DE	370	11	381
Model	EE	30	11524	11554
	Total	400	11535	11935

Estimates of μ_0 and τ are identical to three digits. The γ parameter is equal to 3.61 which indicates that the proximity model increases the prior probability of a marker being associated with a nearby transcript by a factor of about $\exp(3.61) \sim 40$. We now compare the conclusions regarding which transcripts are differentially expressed. One might select the false discovery rate which implies a threshold for the posterior probability. An FDR cutoff of 0.05 is often used, but this can lead to false discoveries that can be easily avoided. The rule that we adopted is that the posterior probability of DE is > 0.95 which implies a FDR of 0.011. The genes selected by this criterion for the posterior probability of DE are very similar in both models as shown in Table 10.

The proximity model selects most of the genes selected by the MOM model as well as 30 other genes which is consistent with higher sensitivity of the proximity model in finding associations between genotypes and gene expression when $\gamma > 0$. It is important to point out that the structure of the proximity model will map more transcripts to the closest marker on the genome. All 30 of the transcripts selected as DE only by the

proximity model were *cis*, and all 11 of the transcripts selected as DE only by the MOM model were *trans*. Some of these transcripts with the largest differences between the two methods are shown in Figure 8.

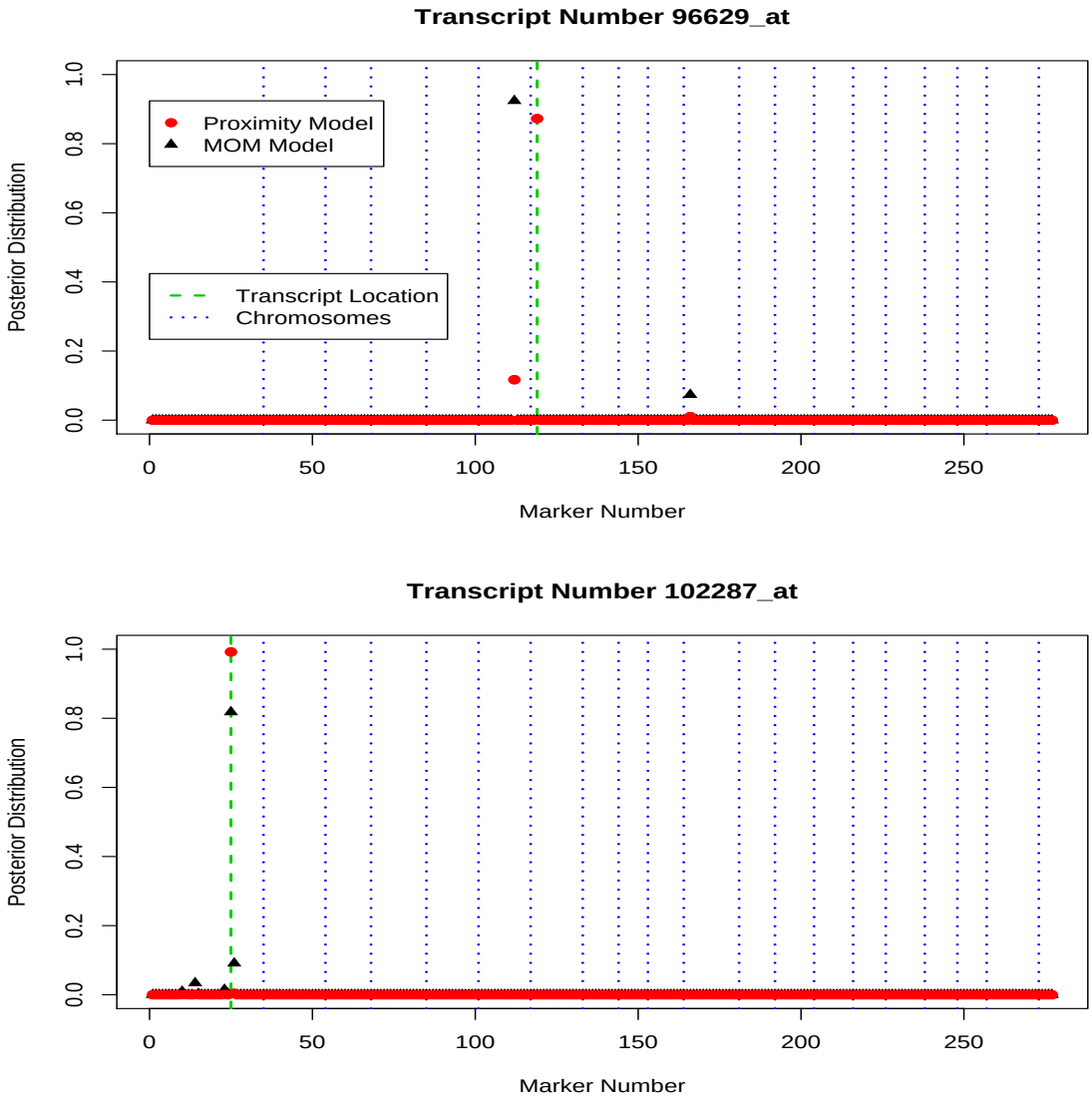

Figure 8: BXD Analysis: Posterior distribution of eQTL for selected transcripts. Transcript 96629_at shows differing posterior distributions between the models. Transcript 102287_at demonstrates the proximity model gives a highly localized eQTL HPD compared to a weaker, more diffuse posterior distribution of the MOM model.

In Figure 8, the posterior modes are mapped more closely to their genomic location by the proximity model while the MOM model may map the transcripts to different

chromosomes resulting in quite different posterior inference. These models were used to select the most likely marker to be associated with the differentially expressed transcripts (m_t^*). Of the 370 transcripts selected by both models, 15 had different values for m_t^*. This would likely affect the results of any forward building model selection process based upon m_t^*. We performed the two eQTL analysis on this dataset, and we found that among the 400 transcripts selected to be associated with one eQTL, there were 185 transcripts declared to be associated with 2 eQTLs with posterior probability greater than 95% given the first eQTL.

We considered the robustness of inference to the selection of the ridge parameter, and values of $\lambda = \frac{1}{2\sigma_\beta^2}$ ranging from 10^{-2} to 10^{-4} did not greatly effect inferences regarding transcripts. Further, the selection of the number of discrete variance points from 10 to 30 did not greatly affect inferences.

3.9 Discussion

We have developed a proximity model for finding eQTL which includes explicitly modeling the effect of genome location by extending the MOM model of Kendziorski et al. (2005). The model adds a data-driven increase in the prior probability of a transcript mapping to the marker that is closest on the genome. We have also extended the MOM model to include more than one eQTL and have allowed for transcript specific variances. We have shown that this proximity model can be implemented in an efficient manner, and that it has favorable performance in terms of power while controlling the false discovery rate. The proximity model performs well compared to the MOM model in simulated

datasets with moderate to large proximity effects, but the MOM model may have more power to detect *trans* associations even when the proximity effect is large. The model was applied to an experimental dataset, and the MOM and the proposed model gave similar overall results regarding the genetic control of transcripts. However, the proximity model suggests that a greater number of transcripts are associated with eQTLs than the MOM model, and it was shown that the MOM model gave different localizations of eQTLs regarding some transcripts. Furthermore, we showed that roughly half of the transcripts are likely to be associated with a second eQTL. We chose a simple model for the increase in prior probability of a transcript mapping to the closest marker. However, we considered more complex models such as ones that would make the prior probability a decreasing function of the distance of a marker and the transcript. These other models add interpretational and computational difficulties, and they do not reflect any widely accepted biological understanding not captured by the model we used. Lastly, we note that controlling the FDR alone could result in poor decision rules for certain cases as the FDR is a summary statistic for a set of hypotheses, but does not strictly control the posterior probability of the individual alternative hypotheses. Thus, we suggest controlling the posterior probability of individual hypotheses instead.

In future research, we will explore how the ridge parameter may be estimated from the data through a mixed model or empirical Bayes methods which would use the information across markers to regularize the parameter estimation. We also note that this framework of modeling the prior probability of differential expression across transcripts and markers can be readily extended to include other parameters for sequence characteristics like the presence of features in the sequences known to involve transcription regulation. We chose

the prior for the error variance π_{σ^2} to be discrete and noninformative, but future methods could model this density semiparametrically so that the prior becomes more informative resulting in higher power. Also, we used a limited number of imputations of missing genotypes, but in applications involving less dense maps, more imputations should be performed which might require more efficient computational implementations.

4 Microarrays for Binding Site Discovery

The third paper advances an analytic method that integrates microarray data and the sequence analyses of the probes to discover transcription factor binding sites. The specific hybridization of DNA microarrays can be used in ways other than the measurement of expression. The method of Chromatin (Ch) Immunoprecipitation (IP) microarrays (ChIP-Chips) use microarrays of DNA sequences to measure specific DNA-protein interactions. The ChIP-chip technique is reviewed by Buck and Lieb (2004). The goal is to discover the genomic locations of transcription factor binding sites (TFBSs). Transcription factors are proteins that regulate the expression of nearby genes by binding to DNA and interacting with RNA polymerase (Stryer, 1995) which is the enzyme responsible for transcription. Gene transcription is often locally regulated so that knowing the location of the TFBS can give insight into which genes are regulated by the transcription factor. Identifying the conditions that transcription factors are active in is important to understanding the role a transcription factor has in certain biological processes and developmental stages. Transcription factors bind to TFBSs with specific sequence patterns that are usually on the order of 10 nucleotides in length, and even in relatively small genomes, the binding sites occur in thousands of locations (Buck and Lieb, 2004). The chromatin immunoprecipitation procedure concentrates specific DNA-Transcription Factor complexes in the following manner. First, the transcription factor of interest binds to the DNA *in vivo* under controlled conditions, and the extracted protein-DNA complexes

are fixed or crosslinked. The DNA is broken into 1kb fragments by sonication. Next, an antibody specific to the transcription factor of interest binds to the protein-DNA complex, and this entire complex precipitates out of solution. The DNA is then extracted, and the DNA is amplified and labeled. We call this the IP sample. Control samples of DNA that do not go through the IP process are used as a reference, and either two color microarrays (Buck and Lieb, 2004) or high density oligonucleotide arrays (Kapranov et al., 2002; Cawley et al., 2004) compare the DNA present in the IP sample and the control at each locus. The ChIP process is shown in Figure 9.

If a locus or continuous region of many loci has higher intensity in the IP sample than the control, it is said to be *enriched*. The sequences of the enriched regions are analyzed for the presence of a *motif* or specific TFBS corresponding to the transcription factor of interest. Current methods have separated ChIP-chip analysis into two steps that first identify enriched regions then estimate the TFBS motif given those regions with a separate procedure. We propose a method that unifies the ChIP-chip and sequence analyses to more accurately estimate the enrichment probabilities and the location of the TFBSs.

4.1 The Data

There are two main technologies used for ChIP-chip experiments. First, there is a two-color system in which the IP sample is labeled with one fluorescent dye and the control sample is labeled with a different dye and applied to the same array. The probes on the two-color arrays range from about 100 to 2,000 base pairs in length. For each probe p on

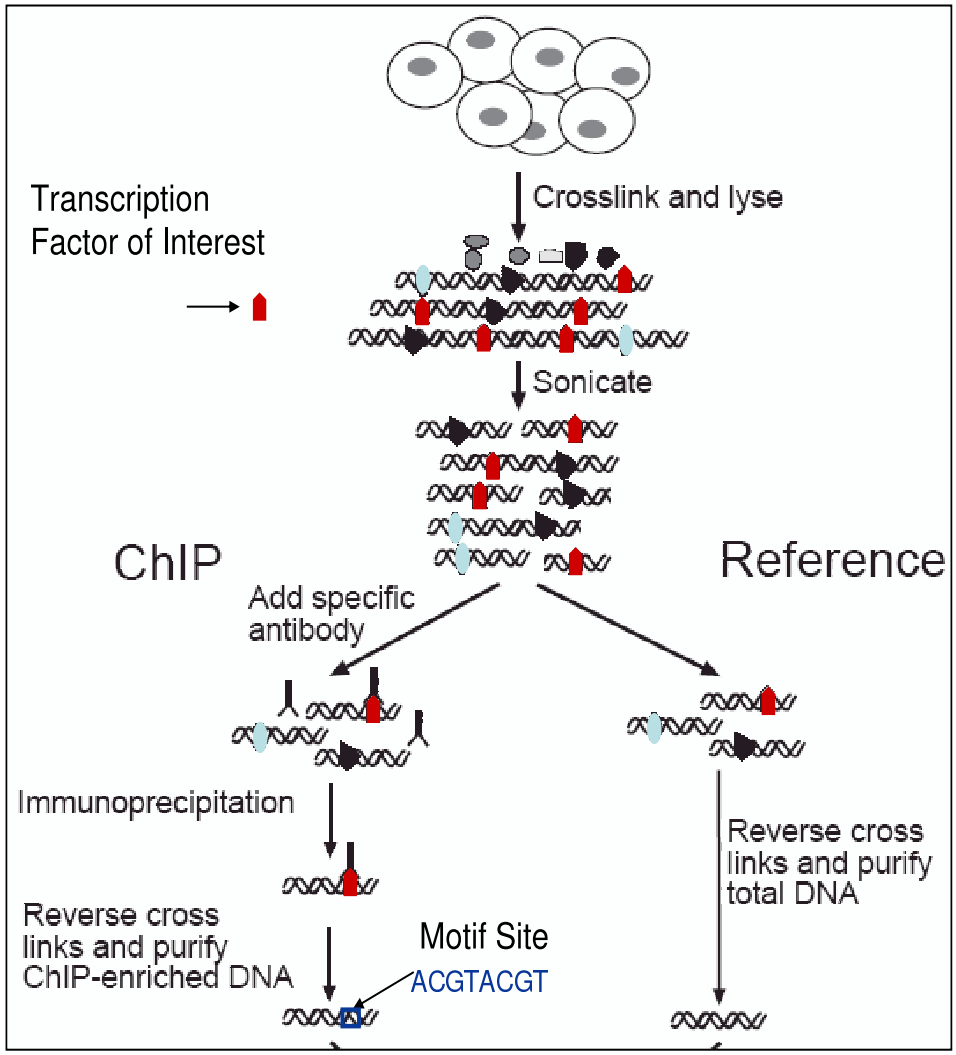

Figure 9: Chromatin Immunoprecipitation Process as shown in Buck and Lieb (2004).

each array, there are two measurements: one for the IP sample intensity (Dye 1) IP_p and one for the control sample intensity (Dye 2) Control_p. The variation due to the random error of a specific probe's measurement is reduced by taking the ratio of $\frac{\text{IP}_p}{\text{Control}_p}$ which removes the multiplicative effect of probe p that is common to both IP_p and Control_p (Rocke and Durbin, 2001). Enrichment implies that $\log(\frac{\text{IP}_p}{\text{Control}_p}) > 0$ for a given probe p. The second type of ChIP-chip is the oligonucleotide array. Oligonucleotide arrays have probes that are 15-30 base pairs in length and have only one fluorescent sample applied to each array. For oligonucleotide arrays, the IP sample is applied to one array or set of arrays, and the control sample is applied to a different array or set of arrays. Enrichment implies that $\text{IP}_p > \text{Control}_p$ where the two measurements have random errors which are uncorrelated as the probes are on separate and independent arrays. The two-color and oligonucleotide enrichment tests are analogous to the paired and unpaired t-test, respectively.

ChIP-chip analysis should also consider the spatial correlation *between* probes that represent adjacent loci. Probes are correlated if the genomic distance between the probes is less than the length of the DNA fragments in the sample. For example, tiling oligonucleotide arrays have been constructed for human chromosomes 21 and 22 that have an average inter-probe distance of about 35 bp (Kapranov et al., 2002) (Cawley et al., 2004; Keles et al., 2004) whereas the distance between the probe midpoints is larger (200-1500 bp) for two-color arrays (Buck and Lieb, 2004). Correlation between adjacent probes is a prominent feature of the data because the DNA fragments ($\sim$ 1 kb) applied to the arrays may span two or more probes (Buck and Lieb, 2004).

In this paper, we focus on two-color ChIP-chip data. The ChIP-chip experiment can

be represented as an $P \times R$ matrix Y where microarray replicates are indexed $r \in [1 \dots R]$, and the probes are indexed by $p \in [1 \dots P]$. A row of this matrix which contains all measurements from a probe is denoted as Y_p. The number of probes P ranges from $10,000$ to $1,000,000$ in different experiments, and the number of replicates R ranges from 4 to about 10. Y_{pr} is the log-ratio of the IP sample intensity and the control sample intensity so that $Y_{pr} = \log(\text{Red}_{pr}/\text{Green}_{pr})$. A schematic of the data is shown in Figure 10. The ChIP-chip data consists of consecutive measurements of $Y_{1,1}$, $Y_{2,1}$, $Y_{3,1}$, etc. The values of Y_{pr} that are higher are more likely to be IP enriched. The histogram of average values of Y_p from the yeast dataset discussed in this paper (Lieb et al., 2001) are shown in Figure 11. The averages can be thought of as a mixture of the enriched and the not enriched probes, and the proposed model density estimates for these two components is shown.

The sequence that corresponds to probe p will be denoted as X_p. The consecutive probes are complementary to adjacent segments on the genome, and the fragments which hybridize to the probes correspond to the surrounding sequence. The genome consists of complementary double helices of DNA, so that each segment of the genome has two sequences which are the reverse complement to one another. X_p consists of these two sequences of A's, C's, G's, and T's with length K_p. The probe sequences can range from a few hundred to several thousand base pairs in length, but the resolution of each probe is limited by the size of the applied DNA fragments (about 1000 bp). A subsequence of X_p from position j to position k will be denoted as $X_p[j : k]$.

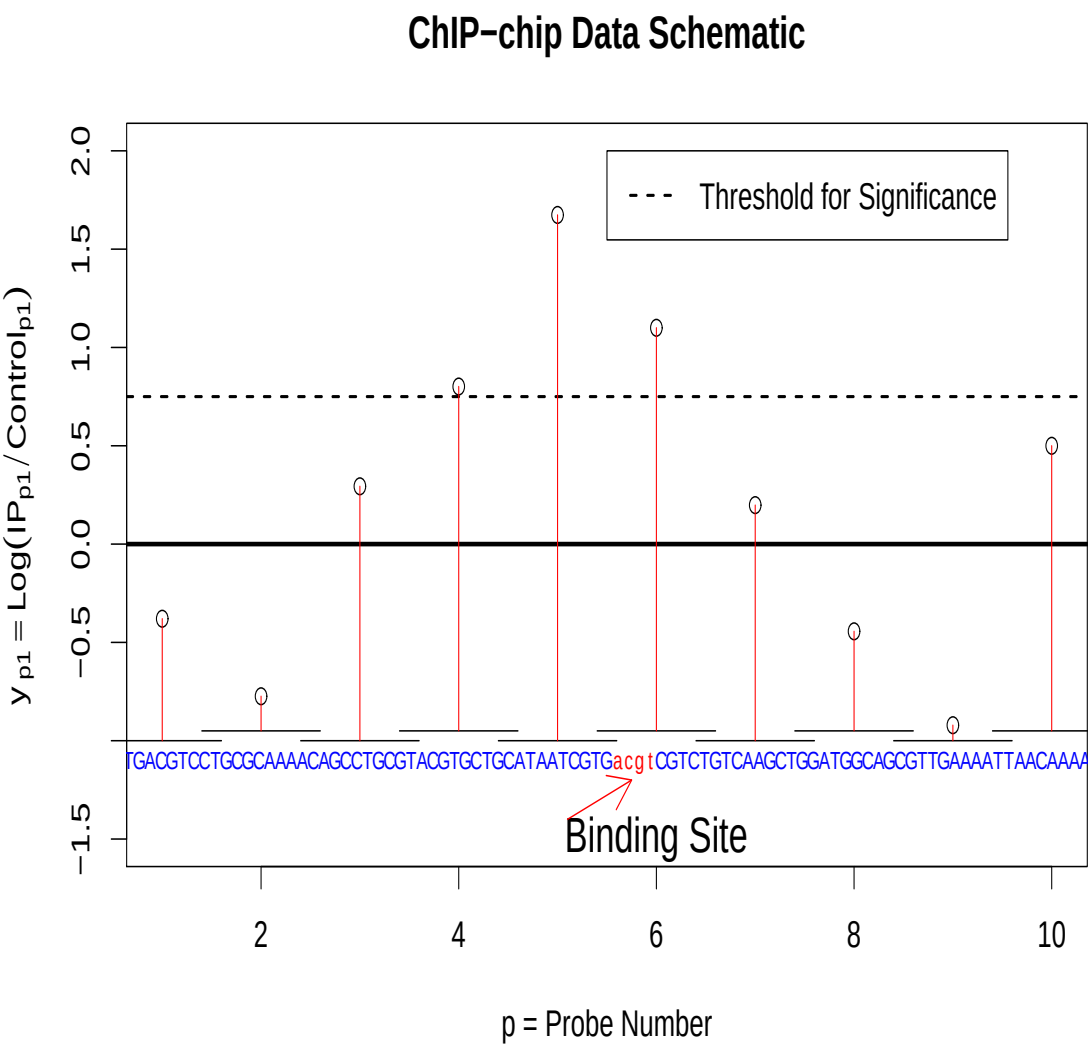

Figure 10: ChIP-chip data schematic is shown for one ChIP-chip replicate. The genomic sequence is shown in blue, and the segments corresponding to the probes is indicated by bars over the sequence. The number of base pairs has been greatly reduced for clarity. Note that $\log(\frac{\mathrm{IP}_p}{\mathrm{Control}_p})$ is increased for the probes close to a binding site, and the region corresponding to the significant probes contains a binding site. Also, note the correlation between adjacent probes.

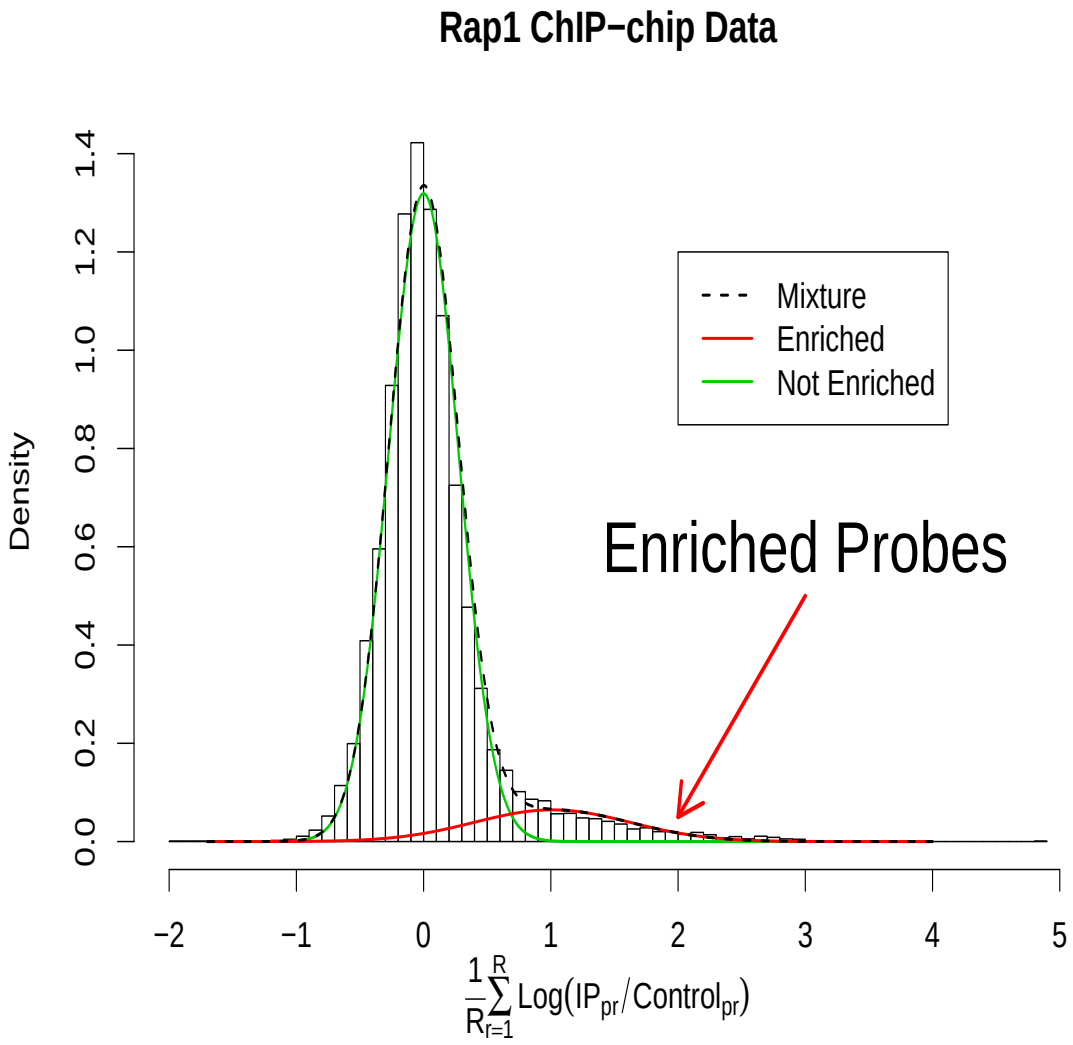

Figure 11: Histogram of average probe intensities $\bar{Y}_p = \frac{1}{R}\sum_{r=1}^{R} Y_{pr}$ from Rap1 yeast experiment. The density estimates from the proposed model fit are overlayed, and the two component mixture of both Enriched and not Enriched probes is evident.

4.2 Current Methods for ChIP-Chip Data

The analysis of ChIP-chip data has been done in two phases. The first phase deals with the microarray data, and it analyzes the intensity to find the regions of enrichment. The second phase uses the sequences of the regions found in the first phase to find the motif. The first analytic phase is discussed below.

4.2.1 ChIP-Chip Analysis to Identify Enriched Regions

The microarray phase of the analysis should consider the spatial correlation between probes that represent adjacent loci. This correlation occurs when the genomic distance between the probes is less than the length of the DNA fragments in the sample. There are a number of methods developed that account for this data feature. First, a sliding window approach has been suggested by Cawley et al. (2004); Keles et al. (2004); Ji and Wong (2005). The sliding window methods average the test statistic over adjacent probes. Cawley et al. (2004) propose using a Wilcoxon rank sum statistic for each probe, Keles et al. (2004) used Welch t-statistic, and Ji and Wong (2005) used both a t-like-statistic which has a shrunken variance estimate and a statistic similar to a posterior probability. These methods identify regions or peaks of intensity as IP enriched when the moving average of the statistic exceeds a threshold, and should give an FDR or posterior probability of enrichment for each region. Cawley et al. (2004) only used a strict p-value cutoff without estimating the FDR. Keles et al. (2004) used a nested-Bonferroni procedure to estimate the FDR, and suggested cross-validation approach for choosing the size of the window.

Ji and Wong (2005) developed a nonparametric method called Unbalanced Mixture Subtraction (UMS) to estimate the posterior probability and FDR for enrichment. UMS is a method for estimating a mixture distribution of two components say $h(t) = \pi_0 f_0(t) + (1-\pi_0) f_1(t)$. The parameter π_0 represents the portion of probes or probe moving averages that are not enriched, and the densities $f_0(\cdot)$ and $f_1(\cdot)$ correspond to the statistics for the not enriched and enriched states respectively. UMS requires that one can obtain estimates for similar mixture densities where one density ($g_0(t) = p_0 f_0(t) + (1-p_0) f_1(t)$) has a higher contribution from the $f_0(t)$ density ($p > \pi_0$), and the other density ($g_1(t) = q_0 f_0(t) + (1-q_0) f_1(t)$) has a higher contribution from the $f_1(t)$ density ($q_0 < \pi_0$). Another condition for UMS is that $f_0(t) \approx g_0(t)$ or that $p_0 \approx 1$. Ji and Wong find sets of adjacent probes to estimate $g_0(\cdot)$ and $g_1(\cdot)$ by using a cutoff for the test statistic. Given these densities and the assumption that $f_0(t)/f_1(t) \to \infty$ as $t \to t_0$ for some t_0, the authors show that the component densities $f_0(\cdot)$ and $f_1(\cdot)$ as well as the mixing proportion π_0 can be estimated, and from this, the posterior probabilities for enrichment are computed.

Another approach to finding regions of enrichment is to use Hidden Markov Models (HMM). HMMs are Markov random processes with latent states that emit random variables whose distributions depend on the state. The HMM naturally incorporates the spatial dependency in the data because the state of the preceding probe affects the state of the next probe. The essential components of this HMM are the three states (start, enriched and not enriched), the parameters describing transition probabilities between states (π_0, τ_0, and τ_1), and the emission densities of the probe intensities of the enriched ($f_1(\cdot)$) and not enriched states ($f_0(\cdot)$). Ji and Wong (2005) used the nonparametric UMS method to estimate the emission densities and the transition probabilities from the start

state. The other transition probabilities were derived using approximations, but were not fit by maximizing the HMM likelihood. Li et al. (2005) implemented an HMM with the same state space, but used normally distributed parametric models for the emission densities. The emission from an IP enriched state had a mean that was 2 standard deviations above the control mean, and the transition probabilities for changing between enriched and not enriched states were the same for both states and fixed before fitting the model. Li et al. (2005) and Ji and Wong (2005) both demonstrate the superior performance of the HMM method over moving average models in terms of power for detecting IP enrichment for small sample sizes. The HMMs were not compared against each other. One problem with both of these methods is that the transition probabilities and the emission densities are not estimated simultaneously by the HMM.

Keles (2006) proposed a hierarchical mixture model for detecting regions of IP enrichment. The model is similar to that of the Bayesian method for differential expression of Kendziorski et al. (2003) and closely related to the parametric model of Newton et al. (2004a) in that the probe intensities have mixture distributions which have parametric components corresponding to null and alternative hypotheses. The model divides the genome into N continuous regions R_i ($i \in 1 \ldots N$), and each region contains L_i adjacent probes. The array data consists of the control j^{th} probe intensity ($j \in 1 \ldots L_i$) for the i^{th} region $X_j(i)$ and the corresponding IP sample intensity $Y_j(i)$. The latent means of $X_j(i)$ and $Y_j(i)$ are $\mu_{j1}(i)$ and $\mu_{j2}(i)$ respectively. This model estimates the posterior probability that a region contains one and only one subregion of IP enrichment or *peak* as an expectation of a latent indicator variable R_i. The existence of the peak ($R_i = 1$) implies that $\mu_{j1}(i) \leq \mu_{j2}(i)$, and $\mu_{j1}(i) = \mu_{j2}(i)$ if there is no peak ($R_i = 0$). The bound-

aries of the peaks have posterior distributions that are estimated by the model. The start position is given by Z_i and the end position is V_i. Z_i and V_i are discrete random variables that take values corresponding to the probes in the region, and w_i is the random variable for the length of the peak in terms of numbers of spanned probes so that $w_i = Z_i - V_i + 1$. The posterior distribution of Z_i and V_i for enriched regions is the important indication of the localization of enrichment within a region. Keles proposes a parametric density for the latent means of the probe intensities, and uses conjugate gamma distributions for the probe intensities themselves. The conjugate intensity distributions are vital to the model because the latent means of the probes are integrated out analytically. The EM algorithm is used to fit the model, and Keles reports that there may be multiple stationary points. Keles also shows that the estimates for the FDR may be anticonservative when the error model is grossly misspecified. However, the author demonstrates that the hierarchical model has superior power for small sample sizes compared to sliding window methods.

The previous methods for ChIP-chip analysis identify the enriched regions of the genome, but the next analytical phase in the two step approach examines these enriched regions and estimates the motif patterns within them. These methods are discussed in the next section.

4.2.2 Sequence Analysis of Enriched Regions

The TFBS discovery within the sequences of the enriched regions is statistically challenging for many reasons. A transcription factor binding motif is not an exact sequence, and it is usually represented by a $4 \times w$ *position specific weight matrix* (PSWM) Θ that

defines a product multinomial distribution where the four rows represent the nucleotides A, C, G and T and the w columns represent the w motif positions (Liu et al., 1995). The element Θ_{ij} is the probability that the nucleotide at position j of the sequence is i, $i = \{A, C, G, T\}$. Searching for patterns of several base pairs within segments of DNA that might be several thousand base pairs long can lead to many false positive matches because there are thousands of potentially similar sites within a single DNA segment. This multiplicity greatly increases the computational burden, especially if many different DNA segments are considered simultaneously. Further, the "background" DNA sequence that does not contain binding sites generally has a highly non-random distribution of nucleotides, for instance, it may contain dependencies between consecutive base pairs, and these patterns can mimic motifs. The computational and statistical challenges of motif discovery have led to the development of a number of statistical model-based methods for motif discovery (Bailey and Elkan, 1994; Lawrence et al., 1993; Liu et al., 1995; Gupta and Liu, 2003; Thompson et al., 2004; Zhou and Liu, 2004; Gupta and Liu, 2005; Shida, 2006) as well as computationally fast and partially heuristic methods (Liu et al., 2001, 2002; Buhler and Tompa, 2002; Keles et al., 2002; Sinha et al., 2004; Elemento and Tavazoie, 2005).

4.2.3 Motivation for a Unified Model

The key difficulties of the two step approach are displayed in Table 11. The first stage categorizes the probes into the columns (Enriched/Not Enriched), and the second stage categorizes the probes sequence into the rows (Binding Site/No Binding Site). The first stage might be considered as a screening test for probes, whereas the second stage is

Table 11: **Different Possible Probe Outcomes**

		Probe Intensity	
		IP Enriched	Not Enriched
Probe Sequence	Binding Site	True Positives	Ambiguous False Neg?
Contains	No Binding Site	Ambiguous False Pos?	True Negatives

similar to a confirmatory test for a probe containing a TFBS. In this case, the diagonal quadrants of the table would consist of the probes that were accurately classified by the first stage where the left upper quadrant represents the true positives, and the right lower quadrant the true negatives, although only the probe sequences classified as enriched typically are searched for a TFBS. The off-diagonals represent ambiguous states in which probes could have been falsely identified as either enriched or not enriched. Another possibility for probes falling into the off-diagonal categories is that there is underlying biological complexity not explained by the simple binding model, as discussed further in Section 4.6. Minimizing the number of probes that fall in the off-diagonal quadrants is the primary motivation for a model the considers the probe sequence and the probe intensity measurements simultaneously.

Another theoretical advantage of the joint model is the more accurate estimation of the binding site probabilities. In the two step approach, the first step estimates the enrichment probabilities $P(E)$ and selects sequences based upon this probability. The

second step estimates the binding site probabilities $P(B|E)$ within those sequences taken to be enriched with a completely different model. However, the second step could ignore the uncertainty in selecting the sequence in determining $P(B|E)$. The proposed method estimates the enrichment and the binding sites simultaneously which yields an estimate of the joint probability $P(B \cap E)$ that takes into account both sources of uncertainty.

The organization of the rest of the paper is as follows. Section 4.3 describes the proposed model. Section 4.4 discusses a set of simulation studies, and Section 4.5 contains an analysis of a yeast ChIP-chip experiment for the Rap1 TF (Lieb et al., 2001). Section 4.6 contains a discussion of the overall results and outline future avenues of research.

4.3 The General Model

In this section, we first describe the models used for the probe intensity, the probe sequences, and the joint HMM framework for the probe intensity and sequence data.

4.3.1 Probe Intensity Model

The probe level data is modeled through a hidden Markov model (HMM). HMMs are defined by the latent or hidden states, the emission densities of the states, and the transition probabilities between these states. The hidden states of the HMM at the probe level are the binding states of probe p denoted as s_p where $s_p = 1$ if the p^{th} probe is IP enriched and $s_p = 0$ otherwise. The intensity emission has density $f_{s_p}(Y_p)$ where $f_0(Y_p)$ is the not enriched density and $f_1(Y_p)$ is the enriched density for the p^{th} probe's intensity vector Y_p. We assume a hierarchical model such that the measurements for a probe Y_p will be a vector of replicate observations related by a probe-specific mean μ_p, and the

r^{th} replicates Y_{pr} are conditionally independent given μ_p so that $Y_{pr}|\mu_p, \sigma_a^2 \sim N(\mu_p, \sigma_a^2)$. This density for Y_{pr} is denoted as $f_{obs}(\cdot)$. The normality assumption is justifiable if one considers the raw intensity values to have a gamma distribution that is close to a lognormal distribution so that the log transformation yields an approximately normal random variable. The distribution for μ_p is defined in the next layer of the hierarchy to be $\mu_p|s_p = 0 \sim N(0, \tau_0^2)$ and $\mu_p|s_p = 1 \sim N(\mu_1, \tau_1^2)$. We denote these priors for μ_p as $\pi_{s_p}(\cdot)$. Figure 11 demonstrates that the observed probe averages $\bar{Y}_p$ may be accurately fit by the proposed mixture normal densities for μ_p. The density for Y_p can be written as $f_{s_p}(Y_p) = \int \prod_{r=1}^{R} f_{obs}(y_{pr}|\mu_p, \sigma_a^2)\pi_{s_p}(\mu_p)d\mu_p$. Integration with respect to the parameter μ_p yields a compound symmetric multivariate Gaussian density for Y_p with mean $s_p\mu_1\mathbf{1}_R$ and covariance matrix $\sigma_a^2 I_R + \tau_{s_p}^2 \mathbf{1}_R\mathbf{1}_R'$ where $\mathbf{1_R}$ is a vector of 1's of length R and I_R is the identity matrix with dimension R.

4.3.2 Sequence Model

The vast majority of the DNA that is not within the binding sites of interest is referred to as the background DNA sequence. Subsequent letters of this background sequence have some dependency on the previous letters. By accounting for dependencies between the adjacent positions of the background sequence, one hopes to more accurately estimate the foreground motif. This dependency is often modeled as having a higher order Markov structure (Liu et al., 2002). However, these Markov models require large numbers of parameters, and some dependencies such as simple repeats are more important than others for distinguishing binding site motifs from background patterns. We propose using PSWMs representing repeats to allow for the modeling of these low complexity back-

ground patterns with fewer parameters than Markov models. Specifically, the PSWMs of the proposed background model include one-letter words (A, C, G, and T) as well as repeats of A's and T's.

Next, we formulate the model for the sequence data in detail. PSWMs in the model will be denoted as $\Theta_v \equiv$ where $v \in [1 \dots V]$, and $\Theta_V \equiv$ is the motif corresponding to the transcription factor of interest. Let Θ_v have length w_v, and let π_v be the prevalence of PSWM v. The emission densities of the sequence are $p_0(X_p)$ and $p_1(X_p)$ for the enriched and not enriched states, respectively. The density $p_0(X_p)$ denotes $p(X_p|\Theta_1, .., \Theta_{V-1})$, the likelihood of observing the sequence X_p given that it was produced by the background set of PSWMs $\Theta_1, .., \Theta_{V-1}$. Similarly, $p_1(X_p) \equiv p(X_p|\Theta_1, .., \Theta_V)$ denotes the same likelihood given that the motif of interest Θ_V could be present in the set of PSWMs generating the sequence. The sets of PSWMs in $p_1(X_p)$ and $p_0(X_p)$ can be considered as *words* that are part of a *stochastic dictionary* (Gupta and Liu, 2003).

The sequence likelihoods $p_0(X_p)$ and $p_1(X_p)$ do not have closed forms and can be calculated using a recursive formula. The likelihood of subsequence $X_p[i:j]$ given that it was emitted from motif Θ_v is denoted as $p(X_p[i:j]|\Theta_v)$ where $j - i + 1 = w_v$ and is given by

$$p(X_p[i - w_v + 1 : i]|\Theta_v) = I[i - w_v + 1 > 0]I[i \leq K] \prod_{j=i-w_v+1}^{i} \prod_{l \in \{A,C,G,T\}} \Theta_{v,lj}^{I[X_p[j]=l]}.$$

The term $I[i - w_v + 1 > 0]I[i \leq K_p]$ makes the probability 0 when the motif would not fit within the sequence. $p_1(X_p)$ are calculated by using a recursive summation involving the terms $\phi_p(k)$ which are the probabilities of X_p up to position k allowing for all possible

motif sites as below

$$\begin{aligned}\phi_p(k) &= p_1(X_p[1:k]) \\ &= \sum_{v=1}^{V} \pi_v p(X_p[k-w_v+1:k]|\Theta_v)\phi_p(k-w_v). \end{aligned} \tag{58}$$

$p_0(X_p)$ is found similarly by allowing $v = 1, .., V-1$.

4.3.3 The HMM Likelihood

Hidden Markov model likelihoods generally cannot be written in a closed form so that a recursive procedure based upon the law of total probability are used in the likelihood computation (Juang and Rabiner, 1991). We use a *forward summation* recursive formula for computing the likelihood of an HMM, described below. We define $g_p(s)$, the forward probability of state s at probe p, as the probability of the sequence of probes up to probe p given that the p^{th} state is s, given by

$$g_p(s) = P(X_p, Y_p|s) \sum_{s_{p-1}\in\{0,1\}} g_{p-1}(s_{p-1})\tau_{s_{p-1},s} \tag{59}$$

where $P(X_p, Y_p|s) = p_s(X_p)f_s(Y_p)$ for $s \in \{0, 1\}$. The states $\{-1, 0, 1\}$ respectively correspond to the *start*, *not enriched*, and *enriched* states. The τ_{ij} parameters represent the transition probability $i \to j$. The likelihood quantities $g_p(s)$ in (59) will be used to draw samples of probe states s_p within the data augmentation method for fitting the model, described further in Section 4.3.5. Prior specification for the joint intensity and sequence HMM is discussed in the following section.

4.3.4 Priors

The priors for the intensity parameters μ_1 were taken to be noninformative ($\propto 1$), and τ_0^2, τ_1^2, σ_a^2 were also take to be noninformative, in other words, $p(\tau_0^2) \propto \tau_0^{-2}$, $p(\tau_1^2) \propto \tau_1^{-2}$, and $p(\sigma_a^2) \propto \sigma_a^{-2}$. The priors for each row of the HMM transition matrix (τ_{ij}) $(i, j = -1, 0, 1)$ are taken to be Dirichlet distributions with hyperparameters denoted as δ_{ij}. More precisely, $[\tau_{si0}, \tau_{si1}] \sim \textit{Dirichlet}(\delta_{i0}, \delta_{i1})$. The δ_{ij} are equal for all transitions so that $\delta_{ij} = \delta_{i'j'}$ and are small (0.1) relative to the total number of transitions $\sim P = 11,575$, and therefore, minimally informative.

One difficulty in estimating the motif is that the motif and prevalence of the motif may be jointly nonidentifiable in practice. The less conserved a motif is, the more prevalent it may be. If there is no prior placed on the motif prevalence, then the model often tends to converge to a highly prevalent and non-specific motif which contradicts the biological understanding of the specificity of transcription factor binding. A relatively strong prior may be implemented for π_V to avoid this problem and hasten convergence. Instead of drawing the PSWM prevalence vector $\boldsymbol{\pi}$ from a Dirichlet distribution, the vector of π_v can be drawn in a hierarchical manner. The prior for the transcription factor motif prevalence is $\pi_V \sim \text{Beta}(\delta_0(1-\gamma), \delta_0\gamma)$ where δ_0 is a large pseudocount and γ (with $0 < \gamma < 1$) indicates the prior expected value. The conditional prior for the other components of $\boldsymbol{\pi}$ $(\pi_1, \ldots, \pi_{V-1})$ can then be drawn from the prior Dirichlet distribution $D(\delta_1, .., \delta_{V-1})$ and scaled by $1 - \pi_V$. $\delta_1, .., \delta_{V-1}$ represent pseudocounts which are set to a small value to avoid Dirichlet parameters of value 0. The prior for the motif matrix of interest Θ_V is taken to be the product Dirichlet distribution $PD(B)$ where B is a $4 \times w_V$ matrix

of pseudocounts where the element B_{ij} is the count of the symbol i at motif position j which is set to a small value to avoid Dirichlet parameters of value 0, but is uniform across letters and not informative. The Data Augmentation (DA) sampling scheme for fitting the full HMM is given in the following section.

4.3.5 MCMC Fitting Procedure

We fit the model with a Data Augmentation (DA) method. The complete steps of the algorithm are given in the Appendix. First, all the model parameters μ_1, τ_1^2, τ_0^2, σ_a^2, Θ_v, $\boldsymbol{\pi}$, τ_{ij}, and $s_1, \ldots s_P$ are initialized. The intensity parameters (μ_1, τ_1^2, τ_0^2, σ_a^2) are sampled using a Metropolis-Hastings (MH) random walk procedure, and the enrichment states s_p are then sampled jointly using the backward sampling technique described in Section 4.3.3. The transition parameters τ_{ij} can be drawn from the complete conditional distribution Beta$(t_{ij} + \delta_{ij}, \sum_{k \neq j} t_{ik} + \delta_{ij})$ where t_{ij} are the $i \to j$ transitions.

Backward sampling generates samples from the joint complete conditional distribution of the vector of s_p. The probability distributions for s_P, $s_{P-1} \ldots s_1$ are given by

$$s_P \sim \text{Bern}\left(\frac{g_P(1)}{g_P(1) + g_P(0)}\right), \tag{60}$$

and for $p \in \{1, \ldots, P-1\}$

$$s_p \sim \text{Bern}\left(\frac{g_p(1)\tau_{1,s_{p+1}}}{g_p(1)\tau_{1,s_{p+1}} + g_p(0)\tau_{0,s_{p+1}}}\right). \tag{61}$$

The s_p imply a segmentation of the entire genomic sequence into enriched regions with $s_p = 1$ and not enriched regions with $s_p = 0$. The enriched segments are formed by the overlapping regions of the genome that correspond to probes with $s_p = 1$, and these

segments are denoted by X_e with length K_e where the index e ($e \in \{1 \dots n_e\}$) stands for Enriched segment.

Another DA algorithm is applied to the X_e in order to sample the motif matrix Θ_V. We define the $K_e \times V$ matrices A_e corresponding to the segments X_e. The elements $A_{e,jv}$ indicate the sampling of the motif or PSWM v at position j such that $A_{e,jv} = 1$ iff the v^{th} PSWM was sampled with $A_{e,jv} = 0$ otherwise. We may sample $A_{e,jv}$ using the backward algorithm described in the Appendix. The motif matrix Θ_V depends on the letter counts from the sampled TFBS where $A_{e,jV} = 1$, and we will call this $4 \times w_V$ count matrix C where the element C_{ij} is the number of the symbol i at motif position j. Θ_V has conditional distribution $PD(B + C)$ where PD is the product Dirichlet distribution. Next, the $\boldsymbol{\pi}$ parameter depends on the number of sampled realizations of each PSWM $n_1, .., n_V$ given by A_e so that $\pi_V \sim \text{Beta}(\delta_0(1-\gamma) + \sum_{v=1}^{V-1} n_v, \delta_0\gamma + n_V)$, and the complete conditional for $[\pi_1, \dots, \pi_{V-1}]$ becomes $\text{Dirichlet}(n_1 + \delta_1, .., n_{V-1} + \delta_{V-1})$.

4.4 Simulation Study

A simulation study was performed to assess the model performance when the true locations of the binding sites are known. The datasets were generated in two ways. First, the sequence was simulated based upon the proposed probe intensity model. Second, the sequence was simulated based upon the TileMap nonparametric ChIP-chip model of Ji and Wong (2005), introduced in Section 4.2.1.

4.4.1 Data Generation

Simulated datasets were generated to assess the operating characteristics of the proposed method compared to the intensity only HMM and the TileMap HMM. The real intensity data was used instead of simulated intensity data in order to mimic the structure and the informativeness of the true experiment. To simulate the sequence data, we used the probe intensity data from the Rap1 dataset (Lieb et al., 2001) with four independent arrays described in Section 4.5, and we applied the intensity only model and the TileMap model which gave the probe enrichment probability estimate $\hat{s}_p$. The enrichment state for each of the probes were then simulated by Bernoulli random variables with probability $\hat{s}_p$, that is, $s_{p,\text{Simulated}} \sim \text{Bernoulli}(\hat{s}_p)$. For the probes that were selected as enriched ($s_{p,\text{Simulated}} = 1$), motif realizations were randomly inserted into the corresponding genomic sequences.

4.4.2 Analysis of Simulated Data

The accuracy of the binding site estimates is used to assess the models. The proposed joint intensity-sequence (IS) model gives the binding site probabilities directly, but the two step ChIP-chip methods like TileMap (TM) give only the enrichment probabilities of the probe sequences, not specific binding sites within those sequences. In order to get binding site estimates, we used the following two step procedure. If a probe sequence had a posterior probability > 0.5 for enrichment, then it was included in the set of selected sequences. These selected sequences were searched for binding sites by fitting the stochastic dictionary model. The primary aim of the analysis is to locate the binding sites of the transcription factor, and these sites may be estimated by the posterior probability

that each position on the genome corresponds to a sampled motif binding site such that $A_{e,jV} = 1$. This probability is estimated by averaging the indicators $A_{e,jV}$ at each position on the genome at each iteration of the DA sampler. A position on the genome was included in a list of binding sites if the posterior probability of being sampled as a TFBS was > 0.5.

When fitting motif discovery models with real DNA used as background, there are multiple motifs that represent multiple modes in the likelihood surface which may result in poor convergence. Multimodality issues when one does not know the true motif are discussed in the Section 4.5. For the simulations, the DA sampler was initialized to the true motif estimate and then updated as per the algorithm. This is done to limit the amount of human supervision that would have been required for de novo motif finding within the many simulated datasets. In the low prevalence scenarios, a strong prior was placed on the prevalence of the motif to prevent divergence as described in the Section 4.3.4 so that $\delta_0 = 10^6$ and $\gamma = 0.0001$. Sensitivity analyses demonstrated that model estimation was robust to prior specifications within a moderately large range of the set values (more details in Section 4.5).

Four models were applied to each dataset. In the first model, the motif sites were sampled with the stochastic dictionary model conditioning upon the true enrichment region. Call this the Known Binding Region (KBR) model. Second, the two step procedure was applied by fitting the Intensity Only (IO) model, and third, the two step procedure was applied using the TileMap method (TM). The TileMap method was not originally designed for two-color arrays, but it is flexible and may use a test statistic for probe enrichment computed by another method. The test statistics for each probe were computed

separately as the p-value under the null hypothesis that $\bar{Y}_p \sim N(0, \hat{\tau}_0^2 + \sigma_p^2/R)$ where $\hat{\tau}_0^2$ is given by the IO model, and σ_p^2 is a shrinkage estimate for the variance suggested by Ji and Wong (2005). Lastly, the proposed joint intensity and sequence (IS) model was applied. The model performance assessment was in the sensitivity and Positive Predictive Value (PPV) for detection of simulated binding sites.

4.4.3 Intensity Only Model Simulations

The first step was to fit the intensity only model to the array data which results in enrichment estimates $\hat{s}_p$ from which the enrichment probes are selected as described in Section 4.4.1. There were four simulation scenarios with two levels of motif conservation, and two levels of motif site prevalence. The two simulated motifs were a highly conserved motif and the Rap1 binding motif taken from the literature. The highly conserved motif consisted of a 13 length sequence with each position having a 99% probability of the consensus letter and the rest of the letters with equal probability. The two levels of motif prevalence were 0.0005 (High) and 0.0002 (Low). Each of the four scenarios was repeated 5 times. The results are shown in Table 12.

As one might expect, the highly conserved motif was detected more accurately than the Rap1 motif for all models. Also, the decreasing the prevalence of the Rap1 motif negatively impacted the sensitivities of all models. However, the effects of motif conservation were the most profound. The IS model was almost equivalent to the KBR model with the artificial motif. The IS model gives superior performance compared with the IO model in terms of both sensitivity and specificity for all four scenarios. Most notably, the PPV is enhanced by the joint IS model for all scenarios from ~ 65 for the IO model

Table 12:
Simulations Based on Intensity Model Enrichment Estimates
KBR = Known Binding Region; IO = Intensity Only Model;
IS = Joint Intensity Sequence Model; TM = TileMap Model

Table 2a - Highly Conserved Motif and Prevalence = 0.0005

Outcome (SD)	True	KBR	IO	TM	IS
Sensitivity	-	91.8 (1.2)	77.7 (1.4)	37.6 (1.2)	90.1 (1.0)
PPV	-	93.5 (6.5)	63.4 (1.3)	63.2 (0.6)	93.2 (0.8)
Total Sites	839 (31)	821 (31)	1016 (57)	505 (23)	829 (30)

Table 2b - Highly Conserved Motif and Prevalence = 0.0002

Outcome (SD)	True	KBR	IO	TM	IS
Sensitivity	-	91.9 (2.0)	78.8 (3.2)	37.5 (1.2)	91.4 (2.6)
PPV	-	93.2 (1.3)	62.5 (1.5)	62.1 (0.9)	92.9 (1.7)
Total Sites	336(21)	331 (22)	423 (29)	209 (13)	331 (24)

Table 2c - Rap1 Motif and Prevalence = 0.0005

Outcome (SD)	True	KBR	IO	TM	IS
Sensitivity	-	70.5 (1.3)	57.7 (1.7)	23.4 (1.8)	63.7 (1.5)
PPV	-	95.7 (0.5)	63.4 (0.9)	66.7 (0.3)	97.2 (1.0)
Total Sites	868(21)	639 (24)	790 (21)	306 (33)	569 (14)

Table 2d - Rap1 Motif and Prevalence = 0.0002

Outcome (SD)	True	KBR	IO	TM	IS
Sensitivity	-	57.2 (4.2)	41.7 (3.0)	7.9 (8.9)	45.5(5.0)
PPV	-	95.3 (2.2)	67.3 (1.8)	33.8 (32.7)	98.2 (1.6)
Total Sites	339 (22)	204 (22)	211 (30)	131 (24)	158 (25)

compared with ~ 95 for the IS model. This implies that the motif matrix estimation is more accurate for the IS because this estimation is directly related to the accuracy of binding site estimation. In the low prevalence Rap1 motif scenario, TileMap procedure failed to find any binding sites in 2 of the 5 simulations.

4.4.4 Simulated Sequence Based on the TileMap Model

Next, we simulated sequences based upon the nonparametric TileMap intensity model enrichment estimates. The TileMap intensity model selected fewer regions to be enriched than the intensity only model, and a higher prevalence of binding sites within these regions was needed in order to estimate the motif accurately. The Rap1 motif was randomly inserted into the selected regions with a prevalence of 0.001. This scenario was repeated 5 times, and the results are shown in Table 13. The TileMap (TM) model has the highest sensitivity of the three models, but the joint IS model is demonstrated to have superior PPV compared to the TM model (98% vs 67%) with little penalty in terms of lost sensitivity (56% vs 64%). The nonparametric intensity model of TileMap assumes that the probe intensity component of the proposed model may be misspecified, but the proposed joint model still shows an excellent performance.

4.5 Yeast Data Case Study

We considered a yeast dataset from Lieb et al. (2001) which was a ChIP-chip experiment for the Rap1 transcription factor.

Table 13:
Simulations Based on TileMap Enrichment Estimates
KBR = Known Binding Region; IO = Intensity Only Model;
IS = Joint Intensity Sequence Model; TM = TileMap Model

Prevalence = 0.001

Outcome (SD)	True	KBR	IO	TM	IS
Sensitivity	-	77.5 (0.9)	55.9 (2.0)	63.9 (2.1)	56.3 (2.2)
PPV	-	93.3 (1.0)	70.0 (1.6)	67.1 (1.0)	98.3 (0.8)
Total Sites	656 (33)	595 (22)	566 (33)	623 (40)	391 (21)

4.5.1 Data Preprocessing and Initialization

The data consist of four arrays and $11,575$ non-telomeric probes of various lengths spanning the yeast genome of 17 chromosomes with a total of 12 million base pairs. Simple repeats were removed from the genome with the RepeatMasker software (Smit et al., 2004). We used median centering and variance standardization to normalize the data. Shifting the median of each array to 0 is important because the proposed model assumes that the majority of the observations will arise from a distribution that is symmetric about 0. The model also assumes that the within array variance is equal which motivates variance standardization. This procedure produced good results as seen later in this section, but there is still a need for improved normalization methods designed for ChIP-chip data (Buck and Lieb, 2004).

We preprocessed the data through an initialization phase that is similar to the first step of the two stage procedure in which the segments of highest enrichment are selected using the intensity only model. This step will allow the model to avoid the multimodality

difficulties of sequence modeling by providing the initial estimate of the motif matrix that will later be updated by the joint model. These probes that were selected by the IO model were ranked according to the log-ratio of the intensity probabilities in favor of enrichment $\log(f_1(Y_p)/f_0(Y_p))$. The sequences of the probes in the highest 1% likelihood ratios were then selected, and the following search for the initial motif estimate was implemented.

The initialization of the sequence model requires a reasonable estimate of the TFBS motif to facilitate convergence. The sequences selected by the above procedure are likely to have the highest concentration of the binding site for the motif, but it is evident that there are many non-random patterns in the DNA that correspond to different modes in the likelihood and can lead to the failure of the stochastic dictionary model to find the motif which gives the highest likelihood for these sequences.

An accumulating stochastic dictionary model was fit to the sequences in which successive motifs are estimated and then added to the dictionary which accumulates these new motifs. First, the dictionary was initialized with PSWM of length one representing A's, C's, G's, and T's as well as repeat words of A's and T's of both of length 4 and length 8. These 8 motifs were considered part of the fixed background model with motif matrices $\Theta_1, \ldots \Theta_8$. The search was restricted to the previously reported motif width of 13 (Lieb et al., 2001), and a motif of length 13 with uniform probability across all letters at all positions was added to the dictionary and updated using the data augmentation method described in the Section 4.3 ($V = 9$). The prevalence of this motif is fixed to be 0.0001, and this motif is considered the foreground motif Θ^* and is the only motif updated in each cycle of the DA sampler. After approximate convergence, the updated motif is added to the fixed background dictionary, and another motif of length 13 with

uniform probability across all letters at all positions is added to the dictionary so that $V = 10$, and this new word becomes the new updated foreground motif. The procedure of iteratively adding words to the background allows the model to consider different modes in the space of potential motifs.

Two likelihoods of the sequences are plotted across the iterations in order to find a reasonable motif for initialization. The first is the likelihood of the sequences given the full dictionary up to that point which may be denoted as $\prod_{X_i \in \text{Top Sequence}} p(X_i|\Theta_1, \ldots \Theta_{8+m}, \Theta^*)$ where $m \geq 1$ is the number of accumulated words and Θ^* is the updated motif. The recursive relationship (58) is used to calculate the likelihood of the stochastic dictionary. The likelihood generally increases as motifs are added to the dictionary, and after a few iterations a plateau is reached signifying entrapment in a likelihood mode. The second likelihood computed is based on the original eight-PSWM background with only the current foreground motif and may be denoted as $\prod_{X_i \in \text{Top Sequence}} p(X_i|\Theta_1, \ldots \Theta_8, \Theta^*)$. This likelihood is an indication of the improvement in model fit given the addition of only the current foreground motif. These two plots are shown in Figure 12 for 4 runs of length 2000 in which a total of 8 words are added to the dictionary every 250 iterations. One can see that runs 2 and 3 have the highest likelihood and that the fifth motifs added in both of the runs give the largest improvement in model fit. The fifth motifs added in both of these runs are very similar. The final estimate for the Rap1 shown in Figure 13 as well as the initialization motif, and there is a strong resemblance to the motifs reported previously by (Lieb et al., 2001) and in the TRANSFAC database (Matys et al., 2003). We conclude that the motif that gives the largest increase in sequence likelihood is a reasonable choice for the initial estimate in the joint sequence and intensity model.

The next phase of the analysis is the application of the joint model. An assessment of the sensitivity to the selection of hyperparameters was performed as well as a comparison of the results with other ChIP-chip methods in the next section.

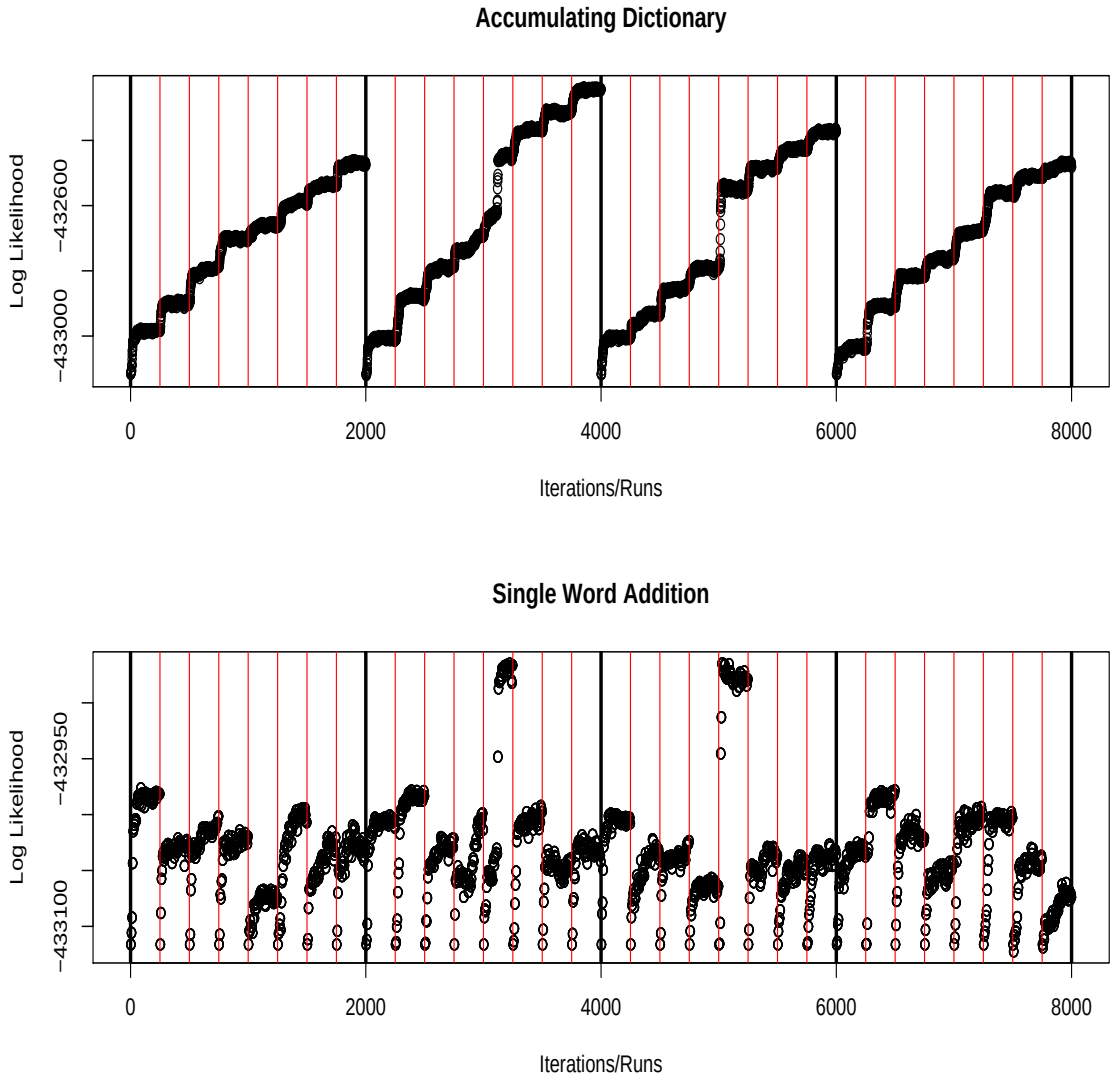

Figure 12: Likelihood trace plots for Accumulating Dictionary (Upper), and Single Word Addition (Lower). Independent runs are distinguished by bold vertical bars, and subsequent motifs are separated by light vertical bars. The Rap1 motif is discovered as the fifth motif in runs 2 and 3.

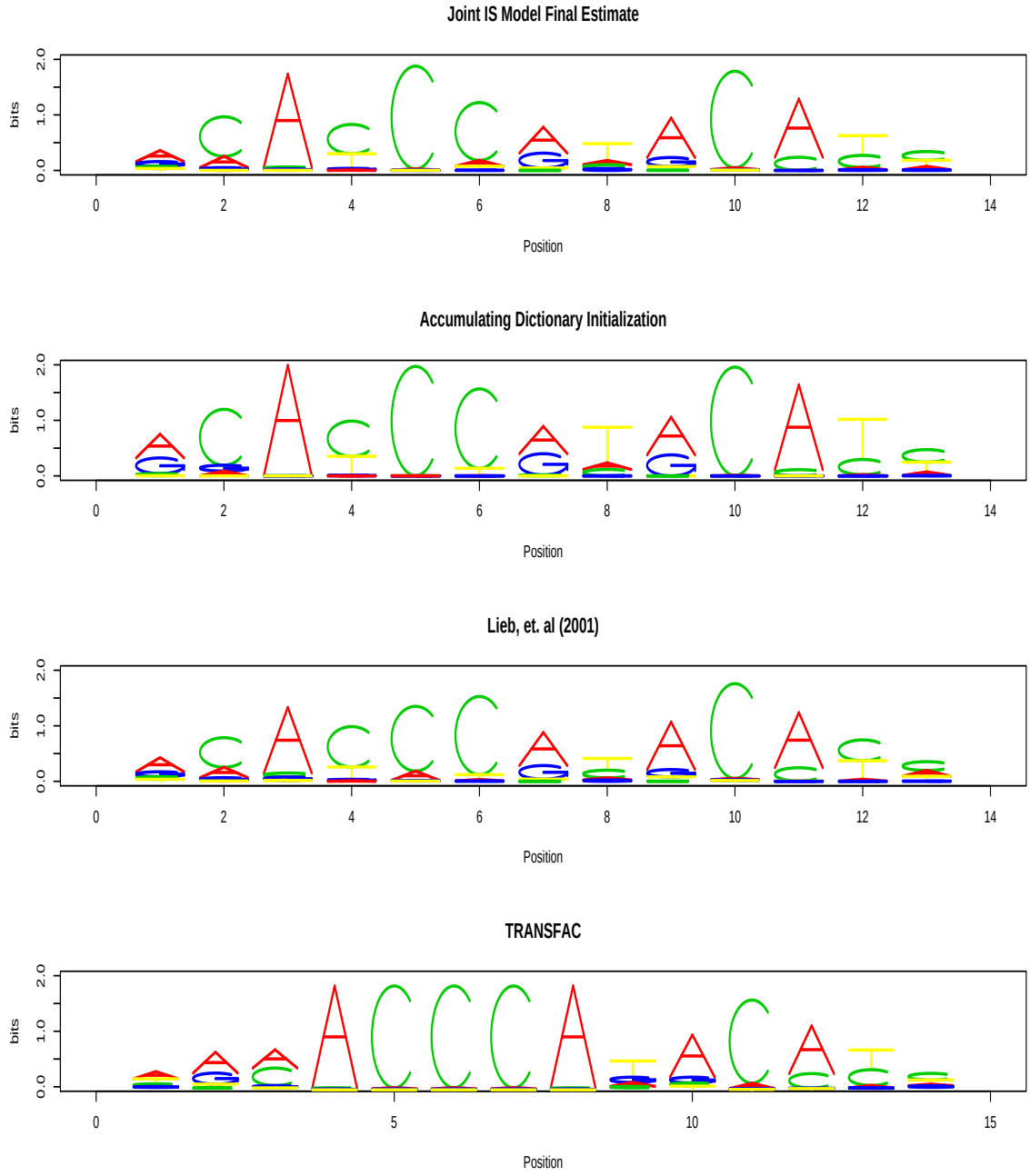

Figure 13: Comparison of motif logos of the model estimates and literature. The final motif estimate for Rap1 by the joint IS model is at the top. Second from the top is the initial estimate given by the accumulating dictionary. The bottom two plots show the motif discovered by Lieb *et al.* (2001) and the motif listed in the TRANSFAC database.

4.5.2 Sensitivity Analysis

The next phase of the analysis is the application of the joint model. An assessment of the sensitivity to the selection of hyperparameters was performed as well as a comparison of the results with other ChIP-chip methods.

We first did a sensitivity analysis to examine the dependence of the final estimates on the choice of the prior hyperparameters. The hyperparameters for the pseudocounts δ_{0ij}, δ_v, and the elements of the pseudocount matrix B were set to 0.1. These pseudocounts are quite small compared to the number of observed counts, and do not greatly affect inferences. We fixed $\delta_0 = 10^6$ and varied the prior parameter for the expected motif prevalence $\gamma \in \{5.0 \times 10^{-5}, 6.0 \times 10^{-5}, 7.0 \times 10^{-5}, 8.0 \times 10^{-5}, 9.0 \times 10^{-5}, 10.0 \times 10^{-5}, 20.0 \times 20^{-5}\}$ to assess the sensitivity to this prior. To initialize the IS model, the IO model DA sampler was repeated for $1,000$ iterations for a burn-in period, and the parameter estimates from the $1,000^{th}$ iterations were used as the initial values for the IS model. The IS model DA sampler was repeated for $1,000$ more iterations, and the last 750 were sampled for posterior inference. MCMC convergence of the DA sampler was diagnosed with parallel chains by using the Gelman and Rubin $\sqrt{\hat{R}}$ statistic. The joint IS model was then applied and the corresponding number of binding sites found for each value of γ were $\{278, 290, 293, 297, 306, 311, > 1000\}$ respectively. The last value indicated that the model did not converge to the correct mode of the posterior distribution. One can see that the number of TFBSs was about 300 in the range $\gamma \in [7.0 \times 10^{-5}, 10.0 \times 10^{-5}]$. The positions of the binding sites discovered were also very consistent, the intersection of the binding site lists for each consecutive value of γ being $\{277, 287, 291, 297, 303, -\}$.

In other words, all 277 of the 278 TFBSs found when $\gamma = 5.0 \times 10^{-5}$ were also found when $\gamma = 6.0 \times 10^{-5}$.

4.5.3 Comparisons with Other Methods

We chose the largest value of $\gamma = 10^{-4}$ for which convergence was observed to compare the IS method with three two step methods. The first method is the intensity only (IO) model which is the proposed method without the sequence component, the second method is the Chipotle method (Buck et al., 2005), and the third method is TileMap (Ji and Wong, 2005). The Chipotle method requires that one choose a normal approximation or a nonparametric model to estimate the p-value for rejection of the "No Enrichment" null hypothesis, and one must decide on a p-value cutoff for selecting regions for the motif finding stage. We chose the normal approximation method and a p-value cutoff of 0.001. There does not seem to be an objective rule for choosing this cutoff, but this conservative value is consistent with our other models.

These three methods were implemented, and they produced estimates of the regions of IP enrichment to which the stochastic dictionary model was applied with $\gamma = 10^{-4}$ and $\delta_0 = 10^6$ to obtain lists of estimated TFBS as in Section 4.4.2. The estimates for the parameters common to the IO and the IS models are shown in Table 14, and these estimates are quite similar for all parameters. The comparisons of estimated TFBS are shown in Table 15. There is marked agreement between the three methods with the IS model finding the most TFBS and the IO model the next to most. However, the TileMap method found roughly half of the TFBS of the other methods. The TFBS found by the joint IS model included 97.5%, 89.9%, and 96.9% of the TFBS found by the IO model,

Table 14:
Parameter Estimates from IO and IS methods

Parameter	Intensity Only		Intensity with Sequence	
	Estimate	(SD)	Estimate	(SD)
μ_1	0.982	(0.03)	1.01	(0.03)
σ_a	0.1172	(0.001)	0.1172	(0.001)
τ_0^2	0.045	(0.001)	0.045	(0.001)
τ_1^2	0.364	(0.021)	0.346	(0.020)
τ_{00}	0.97	(0.002)	0.97	(0.002)
τ_{11}	0.71	(0.02)	0.70	(0.02)

Chipotle, and TileMap respectively. Also, the IS model was highly consistent in that it found a much larger number of sites compared to TileMap, for example, that found only 52.8% of the Chipotle sites. This might indicate a higher sensitivity of the IS model, but the higher specificity cannot be directly assessed because the locations of all "true" binding sites are not known.

An analysis of the differences between the probe enrichment probabilities estimated by the IO model and the joint IS model was performed to examine the effect of adding the sequence component to the model. Figure 14 shows the posterior probability $P(s_p = 1|D)$ of probe enrichment under the IO and IS models for all of the probes. Most of the posterior probabilities are close to 0 with a smaller cluster of probabilities close to 1. This polarity of probabilities motivates the cutoff of $P(s_p = 1|D) > 0.5$ for selecting

Table 15: **Estimated Binding Site Comparisons of Four Methods**

-	TileMap	Chipotle	Intensity Only	Intensity with Sequence
TileMap	163			
Chipotle	141	267		
Intensity Only	156	235	287	
Intensity with Sequence	158	240	280	305

probes as enriched. The enrichment probabilities for the probes that were identified as enriched by one model and not the other in quadrants A (IS only) and D (IO only) have IO model enrichment probabilities in the range $[0.049, 0.746]$. These probes are neither definitely enriched nor definitely not enriched according to the IO model. The probes above the diagonal $x = y$ have joint IS model enrichment probabilities that were higher than the enrichment probabilities based upon intensity alone. Most of the probes have smaller enrichment probabilities under the joint IS model. The IO model selected 934 probes as enriched while the IS model selected 922 probes as enriched. Even though fewer probes were identified as enriched by the joint IS model, the IS model found more binding sites which is consistent with the IS model selecting probes that are more likely to correspond to binding sites. These data are also consistent with the idea that including sequence in the model can help to classify some of the probes with ambiguous posterior enrichment probabilities so that more probes corresponding to binding sites are identified as enriched.

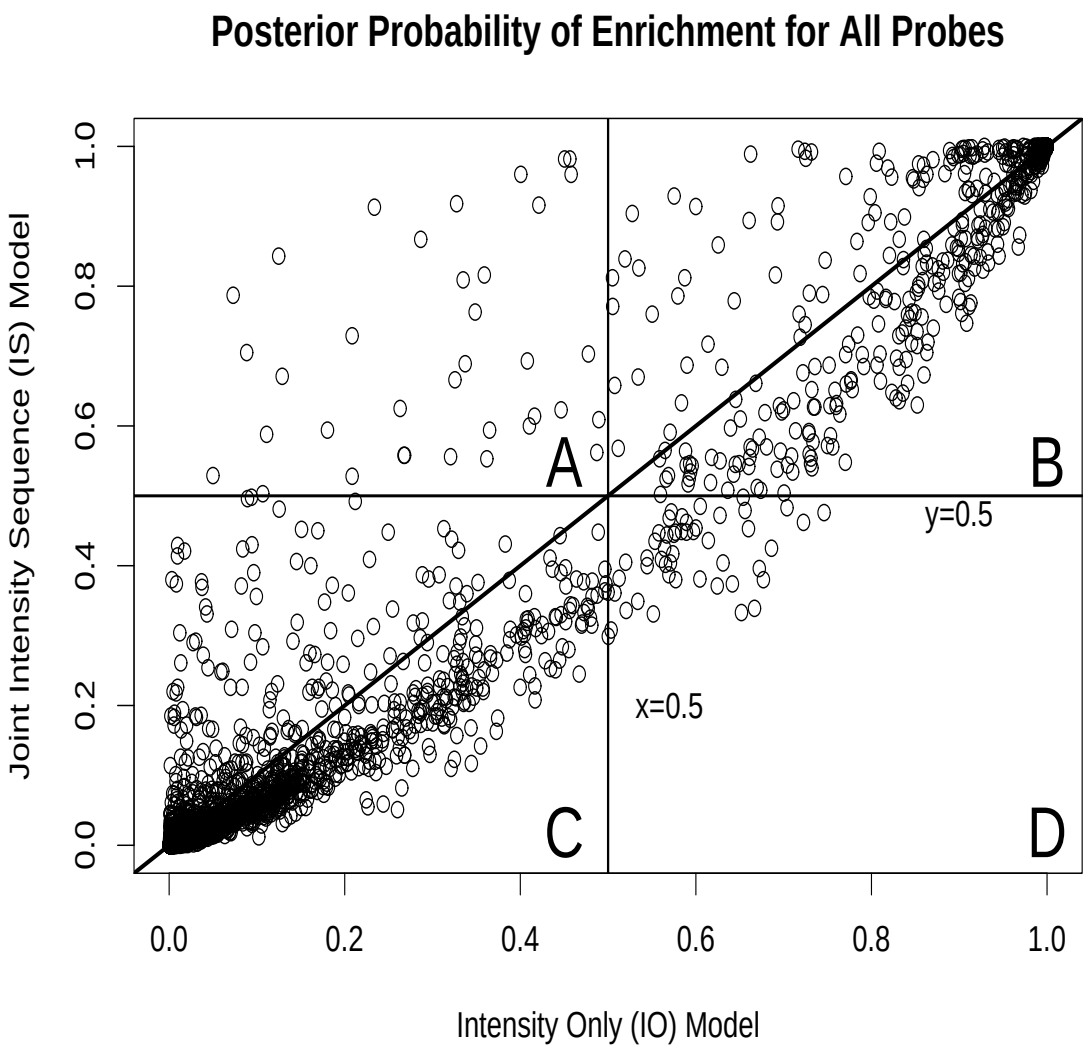

Figure 14: Posterior probabilities for probe enrichment under the Intensity Only (IO) model and Joint (IS) model for all 11,575 probes. The plot has been divided into four quadrants A-D by the lines $x = 0.5$ and $y = 0.5$. The unit line $x = y$ is also drawn. The polarity of the probabilities is evident in the clusters of values at 0 and 1. Probes above the diagonal have enrichment probabilities higher under the IS model, and probes below the diagonal have enrichment probabilities greater under the IO model. The C quadrant contains probes that the IS and the IO model both declared to be not enriched (10,605), and quadrant B contains the probes selected by both models to be enriched (886) The D quadrant contains those probes identified as enriched by the IO model, but not the joint Intensity Sequence (IS) model (48). The A quadrant contains those probes identified as enriched by the IS, but not the IO model (36).

4.6 Discussion

The proposed HMM for transcription binding site detection in ChIP-chip experiments was motivated by the HMMs of Ji and Wong (2005) and Li et al. (2005) with the important extension of jointly analyzing the sequence data rather than the implementation of a two stage procedure. A sequence likelihood based on a stochastic dictionary model is included the emission densities of the HMM. The joint Intensity Sequence (IS) model was shown to significantly out-perform the two stage procedure for binding site discovery in terms of the sensitivity and especially the specificity in the simulated data. The IS model was also applied to a yeast dataset which examined the DNA binding of the Rap1 transcription factor. We proposed a method for overcoming the multimodality difficulties of *de novo* motif discovery using an accumulating stochastic dictionary and choosing the motif that gave the greatest increase in the sequence likelihood. The resulting de novo motif estimate is in close agreement with the Rap1 motif found in the literature. The binding sites estimated by the proposed method from the experimental data were compared to the the binding estimates of the intensity only model, the Chipotle method, and the TileMap method. It is important to note that these three methods do not yield the binding sites directly, but the stochastic dictionary model had to be applied in the second stage. This shortcoming of intensity only models strains comparisons with the IS method because these comparisons would overlook the additional utility of the IS indicating the posterior probability of an *exact* position (1 base pair) on the genome of a binding site rather than merely estimating the posterior probability that a *region* (100-2000 base pairs) might contain one or more binding sites. This utility of the joint model may be considered a

large improvement in resolution in binding site discovery. Nevertheless, the binding sites found by the IS and the two stage approaches were mostly identical, but the IS model estimated the largest number of binding sites. Simulation studies indicated that the joint IS model can successfully estimate binding site probabilities with much higher specificity than two step ChIP-chip analyses that might not accurately combine the uncertainty of enrichment region selection and the uncertainty of binding site identification.

Future work would include several possible variations and extensions of the IS model. First, the IS model currently assumes that each of the large and possibly overlapping sequences corresponding to the probe measurements are either enriched or not enriched. However, with the increasing availability of data at higher resolution, an alternative method might consider modeling smaller non-overlapping segments of DNA having a latent enriched or not enriched state. This might allow for higher resolution, and is another way of pooling the intensity information for adjacent probes. Second, the simple binding model of Table 11 is a reductionist perspective of the transcription factor binding process. The sequence model could be extended to include the possibility of alternative binding motifs for the transcription factor of interest. Alternative motifs could account for some of the probe enrichment not due to the primary motif of interest. Also, the latent state space could be extended from two state (enriched or not) to many states such as "enriched in association with motif 1", "enriched in association with motif 2", etc. Biological insight into transcription factors working in conjunction may also motivate other extensions of the sequence model. As more complex sequence models are developed one might also consider the use of prior information concerning the TFBS motifs.

Appendix

The complete sampling scheme for the joint intensity sequence model is given below.

1. Initialize intensity parameters μ_1, τ_1^2, τ_0^2, and σ_a^2.

2. Initialize sequence parameters Θ_v and $\boldsymbol{\pi}$.

3. Initialize transition parameters τ_{ij}.

4. Initialize probe states $s_1, \ldots s_P$.

5. Sample $\mu_1 \propto \prod_{p=1}^{P} f_{s_p}(Y_p)$ with MH random walk.

6. Sample $\tau_1^2 \propto \prod_{p=1}^{P} f_{s_p}(Y_p)$ with MH random walk.

7. Sample $\tau_0^2 \propto \prod_{p=1}^{P} f_{s_p}(Y_p)$ with MH random walk.

8. Sample $\sigma_a^2 \propto \prod_{p=1}^{P} f_{s_p}(Y_p)$ with MH random walk.

9. Compute $P(Y_p, X_p|s_p) = f_{s_p}(Y_p)p_{s_p}(X_p)$ for $s_p \in \{0, 1\}$.

10. Compute $g_p(0)$ and $g_p(1)$ for $p \in \{1, \ldots, n\}$ with Forward Algorithm.

11. Sample Backwards s_P, $s_{P-1} \ldots s_1$.

12. Count the number transitions t_{ij} where $i \to j$ in $s_{1 \ldots P}$.

13. Sample $\tau_{ij} \sim \text{Beta}(t_{ij} + \delta_{ij}, \sum_{k \neq j} t_{ik} + \delta_{ij})$.

14. Sample $A_{e,jv}$ for all n_e subsequences with algorithm below.

    (a) Initialize $A_{e,jv} = 0$, $n_v = 0$.

(b) Let $j = K_e$ the last position in sequence X_e.

(c) Sample $A_{e,j.} \sim \text{Multinomial}(\frac{\phi_e(j-w_1)\pi_1 p(X_e[j-w_1+1:j]|\Theta_v)}{\phi_e(j)}, .., \frac{\phi_e(j-w_V)\pi_V p(X_e[j-w_V+1:j]|\Theta_v)}{\phi_e(j)})$ so that $A_{e,jv} = 1$ iff the v^{th} PSWM was sampled $A_{e,jv} = 0$ otherwise.

(d) Decrement j by $(w_v - 1)$ iff $A_{e,jv} = 1$.

(e) Increment n_v by 1 iff $A_{e,jv} = 1$.

(f) Return to 3 until $j = 0$.

15. Sample $\Theta_V \sim PD(B + C)$.

16. Sample $\pi_V \sim \text{Beta}(\delta_0(1-\gamma) + \sum_{v=1}^{v=V-1} n_v, \delta_0\gamma + n_V)$.

17. Sample $\boldsymbol{\pi} \sim \text{Dirichlet}(n_1 + \delta_1, .., n_{V-1} + \delta_{V-1})$.

18. Return to 5.

The intensity only sampling scheme would skip steps 2 and steps 14-17, and step 9 would only compute $P(Y_p|s) = f_{s_p}(Y_p)$. The computations of step 9 may be prohibitive because of the terms $p_{s_p}(X_p)$ if the number of background motifs $V - 1$ is large. The ratio $p_0(X_p)/p_1(X_P)$ is what is necessary for the computation of $g_p(s_p)$, and this ratio may be approximated by reducing the number of background motifs in this step.

The MCMC algorithm was implemented using the C laguage, and the applications were run on a Linux cluster with dual-CPU 2.8 Ghz Xeon IBM BladeCenter nodes each with 2.5 GB RAM. The run time for 1,000 iterations of the full model with 10,000 probes with 1,000 bp sequences is approximately 10 hours.

REFERENCES

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62(5):1198–221.

Amos, C., Dawson, D., and Elston, R. (1990). the probabalistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *American Journal of Human Genetics*, 47:842–53.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica*, 4:505–524.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2(4):0511–22.

Bartosiewicz, M., Penn, S., and Buckpitt, A. (2001). Applications of gene arrays in environmental toxicology: fingerprints of gene regulation associated with cadmium chloride, benzo(a)pyrene, and trichloroethylene. *Envirnomental Health Perspect*, 190(1):71–4.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and S., H. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8(8):816–824.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate under dependency. *Ann Stat*, 29:1165–1188.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6:199.

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

Buck, M. J., Nobel, A. B., and Lieb, J. D. (2005). Chipotle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.*, 6(11).

Buhler, J. and Tompa, M. (2002). Finding motifs using random projections. *J. Computational Biol.*, 9(2):225–242.

Burns, T. and El-Deiry, W. (2003). Microarray analysis of p53 target gene expression patterns in the spleen and thymus in response to ionizing radiation. *Cancer Biology and Therapy*, 2(4):444–5.

Butler, N. and Denham, M. (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society, Series B*, 62:585–93.

Byron, S. A., Horwitz, K. B., Richer, J. K., Lange, C. A., Zhang, X., and Yee, D. (2006). Insulin receptor substrates mediate distinct biological responses to insulin-like growth factor receptor activation in breast cancer cells. *British J. Cancer*, 95(9):1220–1228.

Carlborg, O., De Koning, D., Manly, K., Chesler, E., Williams, R., and Haley, C. (2005). *Bioinformatics*, 21(10):2383–2393.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116:499–509.

Chesler, E., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H., Mountz, J., Baldwin, N., Langston, M., Threadgill, D., Manly, K., and Williams, R. (2004). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37(3):233–242.

Chesler, E., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H., Mountz, J., Baldwin, N., Langston, M., Threadgill, D., Manly, K., and Williams, R. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37(3):233–242.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8.

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. (1997). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65 – 73.

Churchill, G. and Doerge, R. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–71.

Davis, S., Schroeder, M., Goldin, L., and Weeks, D. (1996). Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *American Journal of Human Genetics*, 58(4):867–80.

De Paepe, B., Verstraeten, V. L., De Potter, C., Vakaet, L., and Bullock, G. (2001). Growth stimulatory angiotensin ii type-1 receptor is upregulated in breast hyperplasia and in situ carcinoma but not in invasive carcinoma. *Histochemistry and Cell Biology*, 116:247–254.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: P22-37). *Journal of the Royal Statistical Society, Series B: Methodological*, 39:1–22.

Doss, S., Schadt, E. E., Drake, T. A., and Lusis, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research*, 15:681–691.

Dressman, M., Walz, T., Lavedan, C., Barnes, L., Buchholtz, S., Kwon, I., Ellis, M., and Polymeropoulos, M. (2001). Grgenes that co-cluster with estrogen receptor alpha in microarray analysis of breast biopsies. *Pharmacogenomics*, 1:135–141.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarry experiments. *Statistica Sinica*, 12:111–139.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Eisen, M. B., Spellman, P. T., O., B. P., and Botstein, D. (2001). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868.

Elemento, O. and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, 6(2).

Evans, W. and Guy, R. (2004). Gene expression as a drug discovery tool. *Nature Genetics*, 36(3):214–5.

Falconer, D. and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Longman, Harlow, UK.

Fodor, S., Rava, R., Huang, X., Pease, A., Holmes, C., and Adams, C. (1993). Multiplexed biochemical assays with biological chips. *Nature*, 364(6437):555–6.

Fulker, D., Cherny, S., and Cardon, L. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics*, 56(5):1224–33.

Galton, F. (1892). *Hereditary Genius*. Macmillon and Co, London, second edition.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1):1–77.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (Disc: P483-501, 503-511). *Statistical Science*, 7:457–472.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(3):499–517.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46 p541-542 with R. M. Neal). *Applied Statistics*, 44:455–472.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Glonek, G. and Soloman, P. (2003). Discussion of resampling based multiple testing for microarray data analysis by ge, dudoit and speed. *Test*, 12(1):1–77.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Gupta, M. and Liu, J. S. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Statistical Association*, 98:55–66.

Gupta, M. and Liu, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. National Acad. Sciences United States Am.*, 102(20):7079–7084.

Haley, C. and Knot, S. (1992). A simple method for mapping quantitative trait loci in line crosses using flanking makers. *Heredity*, 69:315–24.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., and Brown, P. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):0003.1–21.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York.

Hsieh, W., Chu, T., Wolfinger, R., and Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, 165(2):747–57.

Hubner, N., Wallace, C., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Muller, A., Cook, S., Kurtz, T., Whittaker, J., Pravenec, M., and Aitman, T. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37(3):243–253.

Ibrahim, J., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457):88–99.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, 89:309–319.

Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, 7:805–17.

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

Jansen, R. and Nap, J. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, 17(7):388–91.

Ji, H. K. and Wong, W. H. (2005). Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21:3629–3636.

Juang, B.-H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296:916–919.

Keles, S. (2006). Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, To Appear (Online Early).

Keles, S., van der Laan, M., and Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–1175.

Keles, S., van der Laan, M. J., Dudoit, S., and Cawley, S. E. (2004). Multiple testing methods for chip-chip high density oligonucleotide array data. *Journal of Computational Biology*, 13(3):579–613.

Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A. (to Appear 2005). Statistical methods for expression trait loci (etl) mapping. *Biometrics*.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24):3899–914.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12(1):203–217.

Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, 44:1061–1071.

Koenker, R. and Ng, P. (2003). *SparseM: Sparse Linear Algebra*. R package version 0.61.

Kominsky, S. L., Vali, M., Korz, D., Gabig, T. G., Weitzman, S. A., Argani, P., and Sukumar, S. (2004). Clostridium perfringens enterotoxin elicits rapid and specific cytolysis of breast carcinoma cells mediated through tight junction proteins claudin 3 and 4. *Am. J. Pathology*, 164(5):1627–1633.

Kopreski, M. S., Benko, F. A., and Gocke, C. D. (2001). Circulating RNA as a tumor marker - Detection of 5T4 mRNA in breast and lung cancer patient serum. *Circulating Nucleic Acids In Plasma Or Serum Ii*, 945:172–178.

Lan, H., Stoehr, J., Nadler, S., Schueler, K., Yandell, B., and Attie, A. (2003). Dimension reduction for mapping mrna abundance as quantitative traits. *Genetics*, 164(4):1607–14.

Lander, E. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–99.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214.

Lee, J., Chu, I., Heo, J., Calvisi, D., Sun, Z., Roskams, T., Durnez, A., Demetris, A., and Thorgeirsson, S. (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, 40(3):667–76.

Lee, Y.-J., Mangasarian, O., and Wolberg, W. (2003). Survival-time classification of breast cancer patients. *Computational Optimization and Applications*, 25:151–166.

Lehmann, E. (2005). *Testing statistical hypotheses*. Springer, New York, third edition.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index omputation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31–36.

Li, W., Meyer, C. A., and Liu, X. S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21:I274–I282.

Liao, J., Lin, Y., Selvanayagam, Z., and Shih, W. (2004). A mixture model for estimating the local false discovery rate in dna microarray analysis. *Bioinformatics*, 20(16):2694–704.

Lieb, J. D., Liu, X. L., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics*, 28(4):327–334.

Lipshutz, R. J., Fodor, S., Gingeras, T., and Lockhart, D. (1999). *Nature Genetics, Supplement*, 21:20–24.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J Amer Statist Assoc*, 90:1156–1170.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2001). Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 6:127–38.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20:835–839.

Lobenhofer, E., Cui, X., Bennett, L., Cable, P., Merrick, B., Churchill, G., and Afshari, C. (2004). Exploration of low-dose estrogen effects: identification of no observed transcriptional effect level (notel). *Toxicologic Pathology*, 32(4):482–92.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M, S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–80.

Lynch, M. and Walsh, B. (1998). *Genetic Analysis of Quantitative Traits*. Sinauer Associates, Inc., Massachusetts.

MacAlpine, D. and Bell, S. (2005). A genomic view of eukaryotic dna replication. *Chromosome Research*, 13(3):309–26.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-margoulis, O. V., Kloos, D. U., Land, S., Lewicki-potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). Transfac (R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31:374–378.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278.

Monks, S., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J., Sachs, A., and Schadt, E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics*, 75:1085–1094.

Morley, M., Molony, C., Weber, T., Devlin, J., Ewens, K., Spielman, R., and Cheung, V. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–7.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, 48:829–838.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52.

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

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

Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(4):717–736.

Perou, C. M., Sorlie, T., Elsen, M. B., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S., Lonning, P., Borresen-Dale, A., Brown, P., and Botstein, D. (2000). Molecular portraits of human breast tumors. *Nature*, 406:747–752.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2005). *coda: Output analysis and diagnostics for MCMC*. R package version 0.9-5.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968.

Pounds, S. and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–42.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model (Corr: V71 p219). *Biometrika*, 69:331–342.

R Development Core Team (2004a). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Development Core Team (2004b). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75.

Rhee, M., Wang, Y., Nair, M., and Galivan, J. (1993). Acquisition of resistance to antifolates caused by enhanced gamma-glutamyl hydrolase activity. *Cancer Research*, 53(10 Suppl):2227–30.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8:557–69.

Schadt, E., Monks, S., Drake, T., Lusis, A., Che, N., Colinayo, V., Ruff, T., Milligan, S., Lamb, J., Cavet, G., Linsley, P., Mao, M., Stoughton, R., and Friend, S. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimator and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.

Shen, L. and Tan, E. (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):166–75.

Shida, K. (2006). Gibbsst: a Gibbs sampling method for motif discovery with enhanced resistance to local optima. *BMC Bioinformatics*, 7.

Shultz, V., Phillips, S., Sar, M., Foster, P., and Gaido, K. (2001). Altered gene profiles in fetal rat testes after in utero exposure to di(n-butyl) phthalate. *Toxicological Sciences*, 64(2):233–42.

Sinha, S., Blanchette, M., and Tompa, M. (2004). Phyme: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5.

Smit, A. F. A., Hubley, R., and Green, P. (2004). Repeatmasker open-3.0. *http://www.repeatmasker.org*.

Snyder, A. and Morgan, W. (2004). Gene expression profiling after irradiation: clues to understanding acute and persistent responses? *Cancer and Metastasis Reviews*, 23(3-4):259–38.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the Natinoal Academy of Science U S A*, 98:10869–874.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A., and Liu, E. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population based study. *Proceedings of the Natinoal Academy of Science U S A*, 100:10393–398.

Storey, J., Akey, J., and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3(8):e267.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat*, 31:2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 66(1):187–205.

Stryer, L. (1995). *Biochemistry*. W.H. Freeman and Company, New York, New York, fourth edition.

Tadesse, M. G., Ibrahim, J. G., Gentleman, R., Chiaretti, S., Ritz, J., and Foa, R. (2005). Bayesian error-in-variable survival model for the analysis of genechip arrays. *Biometrics*, 61(2):488–497.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci*, 96(6):2907–12.

Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S., and Lawrence, C. E. (2004). Decoding human regulatory circuits. *Genome Research*, 14(10A):1967–1974.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121.

Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347:1999–2009.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., , and Losick, R. (2004). *molecular biology of the gene*. Cold Spring Laboratory Harbor Press, Cold Spring Harbor, New York, fifth edition.

Wei, J., Greer, B., Westermann, F., Steinberg, S., Son, C., Chen, Q., Whiteford, C., Bilke, S., Krasnoselsky, A., Cenacchi, N., Catchpoole, D., Berthold, F., Schwab, M., and Khan, J. (2005). Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Research*, 65(1):6883–91.

Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing*. Wiley, New York.

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

Yang, Y., Buckley, M., and Speed, T. (2001). Analysis of cDNA microarray images. *Briefings in Bioinformatics*, 2(4):341–9.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Statist Plann Inference*, 82:171–196.

Yvert, G., Brem, R., Whittle, J., Akey, J., Foss, E., Smith, E., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genetics*, 35(1):57–64.

Zeng, Z. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the U.S.A.*, 90(23):10972–6.

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916.