Fast Bayesian Methods for Genetic Mapping Applicable for High-Throughput Datasets

Yu-Ling Chang

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.

Chapel Hill
2008

Approved by

Advisor: Dr. Fred A. Wright

Reader: Dr. Fei Zou

Reader: Dr. Gary Koch

Reader: Dr. Amy Herring

Reader: Dr. Daniel Pomp

Abstract

QTL mapping is a statistical method for detecting possible gene locations (called Quantitative Trait Loci or QTL) and those genes' effects on the variation in a quantitative phenotype, such as the height of a corn plant, etc. QTL mapping has become an important issue in genetic analysis and has made important contributions to the fields of medicine and agriculture. Traditional QTL mapping methods scan the whole genome and calculate the profile likelihood ratios test statistic at each putative QTL location. The maxima of the test statistics for all putative QTL locations are compared with the genome-wide threshold to identify the QTL.

In this thesis, we propose several fast Bayesian methods for QTL mapping, which not only provide direct approximate QTL posterior probabilities at all putative gene locations, but also offer highly interpretable posterior densities for linkage, without the need for Bayes factors in model selection. The applications to simulated data and real data show these methods are highly efficient and more rapid than the alternatives, grid search integration, importance sampling, Markov Chain Monte Carlo (MCMC) sampling or adaptive quadrature. Our results also provide insight into the connection between the profile likelihood ratios test statistic and the posterior probability for linkage. The results of these methods are easy to interpret and have the advantage of producing posterior densities for all model parameters. We infer the presence of QTL at locations with largest posterior probabilities. Because of the high speed and high accuracy of these methods, they are highly suitable for studying high-throughput data sets, e.g. eQTL data sets. The eQTL analysis is a very important application of QTL mapping to a microarray data set, where thousands of transcripts are treated as the phenotypes and provides us insight into the natural variation in gene expression levels. The approach offers highly interpretable direct linkage posterior densities for each transcript, and opens new avenues for research in this area. Biologically attractive priors involving explicit hyperparameters for probabilities of cis-acting and trans-acting QTL are easily incorporated.

We also extend the one QTL Bayesian method to multiple QTL. The advantage of this method is the simultaneous detection of multiple QTL and appropriate modelling of their

joint effects. Multiple QTL mapping can be computationally intensive, even for our efficient Bayesian approaches. Thus, a fully Bayesian multiple QTL approach for high-throughput datasets remains challenging. We investigate a heuristic for conditional search on the two-location search space that shows promise for identifying the global maximum, and offers the potential for extended approximate Bayesian approaches.

ACKNOWLEDGEMENTS

I would like to express my deepest thanks and appreciation to my advisor, Dr. Fred A. Wright for his tremendous help, patience and encouragement during my Ph.D. work. Dr. Wright leads me to this interesting quantitative trait loci statistical analysis field. His sound advice and guidance were invaluable during my research. I have also enjoyed the inspiring discussions with my coadvisor Dr. Fei Zou who keeps encouraging me to explore different problems and providing many useful comments and resources during my research.

I am extremely grateful for the kind assistance, generosity, and advice I received from Dr. Gary Koch at different stages of my study. Thanks to his help and guidance, I can complete my graduate study for both Master and Ph.D. degrees. I would also like to thank him for giving me the opportunity to work in the Biometric Consulting Lab. During my stay in the lab, I have learned so much from the projects, which I participate in. I want to express my thanks to all members in the lab for encouraging and supporting me during my studies.

I am grateful to my committee members for investing time and energy discussing ideas with me. I have benefited greatly from their advice. I thank Dr. Amy Herring for carefully reviewing my dissertation and providing statistical suggestions in Bayesian perspective. I also would like to express my thanks to Dr. Daniel Pomp for providing helpful statistical references and giving useful suggestions from Biological perspective.

This dissertation is dedicated to all my family members for their support and encouragement.

CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

In this thesis, we are addressing the problem of detecting the locations and effects of genes which contribute to the variation of some phenotype. This is called QTL mapping. Moreover, we also expand the idea of QTL mapping to microarray data and gain insight into the effect of variations in gene expression levels. This is called eQTL mapping. Most traditional QTL methods use the log_{10} of profile likelihood ratios test statistic (also called the LOD score in the genetics field) to detect QTL. With these methods, a LOD score for every putative location is computed. We infer the presence of QTL at locations where the LOD scores are above some pre-specified threshold.

In most Bayesian QTL methods, it is customary to draw samples of nuisance parameters from the posterior distributions by applying Monte Carlo sampling methods and then obtain estimates for the unknown parameters based on averaging the samples drawn from these posterior distributions. In our Bayesian method for detecting one QTL, we propose to use the Laplace approximation for the integration of the likelihood function with respect to the nuisance parameters, assuming that the priors of the nuisance parameters are properly uniform distributed. This method is very fast and accurate compared to all other existing Bayesian methods. Thus, our Bayesian method is suitable for high-throughput applications such as eQTL studies. We expand this Bayesian method to detect multiple QTL simultaneously and further propose the iterative Bayesian method via the Laplace approximation for the multiple QTL model. This iterative method has been shown to be relatively fast and has very high accuracy for detecting multiple QTL locations.

In Section 1.1, we introduce QTL mapping as well as some commonly used experimental populations (backcross and F_2) in QTL mapping. We also review some existing

likelihood-based QTL methods and Bayesian QTL methods in this section. In Section 1.2, the introduction to microarray data analysis is provided. In Section 1.3, we explain what the expression Quantitative Trait Loci (eQTL) is, and briefly review some recently developed eQTL methods.

1.1 QTL Introduction

The history of QTL mapping can be traced back to Gregor Mendel's study of the shape of the peas. This classical genetics study involved binary traits, in the sense that the phenotype has only two outcomes, i.e. the shape of peas is round or not. However, most natural phenotypes are quantitative, such as heights or yields of crops. This has motivated the statistical study of the distribution of phenotypes while considering the effects of QTL. In the 1920s, the development of the chromosome theory and genetic linkage helped us to understand the effects of genes on phenotypic variation. In the 1990s, biomedical markers such as Restriction Fragment Length Polymorphisms (RFLPs) and microsatellites were discovered. Since then, there have been many articles studying traits on different organisms, such as pigs (Andersson *et al.* (1994)), maize (Beavis *et al.* (1991), Stuber *et al.* (1992)), mice (Berrettini *et al.* (1994)) and tomatoes (deVicente and Tanksley (1993)) based on these linkage maps. By using linkage maps, additional statistical and biological discoveries have been made.

In the following subsections, we will discuss the experiments that produce backcross and F_2 progeny and the statistical models for QTL mapping, and include literature reviews of the existing likelihood-based and Bayesian QTL mapping methods.

1.1.1 QTL Experiments

We focus below on the data from experimental crosses: backcross (BC) population and F_2 intercross population. There are also some experimental crosses, such as double haploid and some types of recombinant inbred strains, but we will not introduce them here. The breeding process of the experimental crosses usually involves choosing two highly divergent parental strains, each of which is homozygous, e.g. if the genotype of the parental strain is AA at some locus, we called it the "high" parental strain; and if the genotype of the parental

strain is aa at this locus, we called it the "low" parental strain. In the following context, we will focus on the genotypes of their progeny at the same locus. By crossing those two parental strains, we can produce F_1 progeny. The F_1 individuals are heterozygous because they receive one chromosome from the high parental strain and the other chromosome from the low parental strain. The chromosome from the high parental strain has genotype A at the locus and the other chromosome has genotype a at the same locus. Thus all the F_1 individuals have the same genotype Aa at this locus. In order to produce the backcross population, F_1 progeny are crossed back to one of their parents, e.g. the high parental strain with AA. The genotype of the backcross progeny could be AA and Aa with the same probability $\frac{1}{2}$. Then two F_1 strains are intercrossed to produce F_2 progeny. The possible genotypes for the F_2 individuals at this locus are AA with probability $\frac{1}{4}$, Aa with probability $\frac{1}{2}$, and aa with probability $\frac{1}{4}$.

When we consider the genotypes at two loci, the chromosomes of the parental strains during meiosis (the formation of the sex cells) may cross over and recombine. This affects the joint distribution of the genotypes at two loci. The probability of recombination r (also called the recombination rate or recombination fraction) is calculated by Haldane's map function here.

$$r = \frac{1}{2}(1 - e^{-2x}). \tag{1.1}$$

In this formula, x is the map distance between two loci, the expected number of crossovers between two loci, and it is described in unit: Morgan (M). The recombination rate r increases from 0 to 0.5, as the map distance between the loci increases from 0 to ∞.

When using Haldane's map function, "no crossover interference" is assumed, which means that for more than two markers, the recombination event between any two of them is independent of the recombination event between any other non-overlapping two markers. Many other map functions have been proposed, for example, the Morgan map function and Kosambi map function (Ott (1991)) are also very popular, and are used under different assumptions in a more complicated biological process.

Suppose that the high parental line (P_1) has genotypes AA and BB at two loci in which we are interested. The other low parental line (P_2) has genotypes aa and bb at the same two loci. Figure 1.1 shows the cross process of backcross progeny and F_2 progeny. It also shows the distribution of genotypes at these two loci, assuming that the recombination rate between the two loci is r. In this Figure, B_1 shows the distribution of the genotypes of the backcross progeny, which are from the crossing of the F_1 population and the high parental strain P_1. B_2 shows the other distribution of the genotypes of the backcross progeny, which are from the crossing of the F_1 population and the low parental strain P_2. The last line shows the distribution of the genotypes of F_2 progeny, which are from the intercross of the F_1 strain. The purpose of the cross process is to increase the genetic variability of the progeny strains and therefore allows us to detect the possible genes for the variation in the quantitative phenotype.

$$
\begin{array}{ccccc}
\mathrm{P}_1 & & \times & & \mathrm{P}_2 \\
\frac{AB}{AB} & & \downarrow & & \frac{ab}{ab} \\
\downarrow & & \mathrm{F}_1 & & \downarrow \\
\downarrow & & \frac{AB}{ab} & & \downarrow \\
AB & \times & \begin{bmatrix} AB & Ab & aB & ab \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} & \times & ab \\
\downarrow & & \downarrow & & \downarrow \\
\mathrm{B}_1 & & \downarrow & & \mathrm{B}_2 \\
\begin{bmatrix} \frac{AB}{AB} & \frac{AB}{Ab} & \frac{AB}{aB} & \frac{AB}{ab} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} & & \downarrow & & \begin{bmatrix} \frac{ab}{AB} & \frac{ab}{Ab} & \frac{ab}{aB} & \frac{ab}{ab} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \\
& & \downarrow & & \\
& & \mathrm{F}_2 & & \\
\multicolumn{5}{c}{\begin{bmatrix} \frac{AB}{AB} & \frac{AB}{Ab} & \frac{Ab}{Ab} & \frac{AB}{aB} & \frac{AB}{ab}/\frac{Ab}{aB} & \frac{Ab}{ab} & \frac{aB}{aB} & \frac{aB}{ab} & \frac{ab}{ab} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2+r^2}{2} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2}{4} \end{bmatrix}}
\end{array}
$$

Figure 1.1: The breeding process of experimental populations and their recombination rate information, Zeng (2000).

1.1.2 QTL Statistical Model

Suppose that y_i $(i = 1, 2, \cdots, n)$ represents the i^{th} individual's phenotype and also assume that the QTL are located somewhere between the markers in our model. We intend to find the locations and effects of QTL given the markers' genotypes, the markers' locations, and the phenotypes of all individuals.

One QTL model:

First, we consider the most simple model: the one QTL model for backcross progeny. The following equation represents how one QTL genotype affects the distribution of phenotypes:

$$y_i = \mu + a \cdot g_i(x^*) + \epsilon_i, \tag{1.2}$$

where μ is the intercept, a is the additive effect of the QTL, x^* signifies the location of the unknown QTL, $g_i(x^*)$ represents the QTL genotype for the i^{th} individual, $g_i(x^*) = 1$ or -1 if the QTL genotype is Aa (heterozygote) or aa (homozygote), and ϵ_i is the environmental variation with a distribution of $N(0, \sigma^2)$.

Similarly, the one QTL linear model of phenotypes y_i for the F_2 population is shown below:

$$y_i = \mu + a \cdot g_i(x^*) + d \cdot (1 - |g_i(x^*)|) + \epsilon_i, \tag{1.3}$$

where a and d are the additive and dominance effects for the QTL, $g_i(x^*)$ equals 1 if the QTL genotype of the i^{th} individual is AA, 0 if it is Aa, and -1 if it is aa, and μ and ϵ_i are defined the same usually as before in the backcross model.

In the above two models, the phenotypes of the individuals follow a mixture normal distribution since $g_i(x^*)$ is unobserved if x^* is not located at one of the marker locations. The variance σ^2 is defined as a constant.

Multiple QTL model

Some phenotypes are affected by more than one QTL, so multiple QTL models are discussed here. We assume that these QTL act additively, and there may be some interactions between

them in our model.

The equations below show two QTL models for the backcross and F_2 populations, respectively, under the condition that two QTL act additively and independently.

For a backcross population:

$$y_i = \mu + a_1 \cdot g_i(x_1^*) + a_2 \cdot g_i(x_2^*) + \epsilon_i, \tag{1.4}$$

where a_1 and a_2 are the additive effects of two QTL, respectively; x_1^* and x_2^* specify the locations of two QTL on the chromosome.

For a F_2 population:

$$y_i = \mu + a_1 \cdot g_i(x_1^*) + a_2 \cdot g_i(x_2^*) + d_1 \cdot (1 - |g_i(x_1^*)|) + d_2 \cdot (1 - |g_i(x_2^*)|) + \epsilon_i, \tag{1.5}$$

where d_1 and d_2 are the dominance effects of two QTL, respectively.

If two QTL exhibit deviation from additivity (i.e. there is an interaction effect between two QTL), called epistasis, the model will become more complicated. The following equation is the two QTL model with the pairwise interaction for a backcross population:

$$y_i = \mu + a_1 \cdot g_i(x_1^*) + a_2 \cdot g_i(x_2^*) + \delta \cdot g_i(x_1^*) \cdot g_i(x_2^*) + \epsilon_i, \tag{1.6}$$

where δ is the epistasis effect between two QTL.

For the k QTL problem, the model can be generally expressed as:

$$y_i = \mu_{g_i1, g_i2, \cdots, g_ik} + \epsilon_i, \tag{1.7}$$

where $g_i1, g_i2, \cdots, g_ik$ are the joint QTL genotypes for the i^{th} individual, $\mu_{g_i1, g_i2, \cdots, g_ik}$ represents the phenotypic mean of y_i if the i^{th} individual has QTL genotypes: $g_i1, g_i2, \cdots, g_ik$, ϵ_i follows $N(0, \sigma^2)$. For the backcross population, the maximum number of unknown parameters is $2^k + 1$ and the maximum number of unknown parameters for F2 population is $3^k + 1$. Thus the model is quite complicated when we consider k QTL.

1.1.3 Likelihood-Based One QTL Methods

The quantitative inheritance was discovered in the 19th century and arise via the segregation of multiple genetic factors, modified by environmental effects. In this section, we will describe some major one QTL likelihood-based methods that have been used since the early 19th century. For each method, we will discuss the main idea, its advantages and its possible disadvantages.

Binary traits were first described by Gregor Mendel through extensive experiments with the breeding of peas; he found that the shape of peas is either round or wrinkled, i.e. it is a binary trait. However, there are also many other traits which exhibit quantitative variation and which require further investigation.

Thoday (1961) addressed the idea of using genetic markers for identifying multiple genes that control the quantitative variation of some phenotypes. This idea are examined experimentally after biochemical markers such as Restriction Fragment Length Polymorphisms (RFLPs) and microsatellites were discovered. The advantages of using biochemical markers to characterize QTL are their phenotypic neutrality, highly polymorphic properties, and their abundance in the genome.

The following methods are based on whole genome analysis:

Analysis of Variance (ANOVA)

Soller *et al.* (1976) used ANOVA for QTL analysis. The phenotypes of individuals are grouped by the genotypes of the markers. Instead of testing the significance of QTL at some putative locus, we compare the group means between two genotypes of the marker. If the QTL is tightly linked to this marker, then grouping phenotypes according to the genotypes of this marker is essentially the same as grouping phenotypes according to the genotypes of the QTL.

Analysis of variance (ANOVA) is a simple and naive method that permits very fast computation. However, there are some drawbacks with this method. First, we can't estimate the precise location of the QTL. ANOVA only shows which marker is closest to the QTL. Second, when the markers are not dense enough, the linkage between a QTL and its closest marker is weak. The power to detect the presence of a QTL is quite small. Third, if we

estimated the QTL effect by the effect of its nearest marker, we would underestimate its effect; see Lander and Bostein (1989).

The Maximum Likelihood Method (MLE)

To avoid the drawbacks of ANOVA, Weller (1986), Weller (1987) and Simpson (1989) considered the difference between QTL and markers. They include the recombination rate r between a QTL and its markers in the model and test every marker one after another to find whether there is a QTL close to any of them. This method works via the following process: taking the backcross population as an example, it assumes that the individuals with QTL genotype AA have phenotypes distributed as $N(\mu_A, \sigma^2)$, and the individuals with QTL genotype Aa have phenotypes distributed as $N(\mu_a, \sigma^2)$.

For those individuals with the genotype AA at the marker which you are testing, the phenotype has the following distribution:

$$y_i \sim (1-r) \times N(\mu_A, \sigma^2) + r \times N(\mu_a, \sigma^2).$$

But if the individuals have the genotype Aa at the test marker, the phenotype follows the distribution below:

$$y_i \sim r \times N(\mu_A, \sigma^2) + (1-r) \times N(\mu_a, \sigma^2).$$

y_i is the phenotype of the i^{th} individual. r is the recombination rate between the QTL and the marker you are testing. The method uses the EM algorithm to find the maximum likelihood estimates (MLE) for the unknown parameters. One way we can test $H_0 : r = \frac{1}{2}$ $vs.$ $H_A : r \neq \frac{1}{2}$ with the LOD score:

$$LOD = -log_{10}\frac{L(\hat{\mu}_A, \hat{\mu}_a, \hat{\sigma}^2, r = \frac{1}{2})}{L(\hat{\mu}_A, \hat{\mu}_a, \hat{\sigma}^2, \hat{r})}.$$

Alternatively, we calculate the LOD score for each marker locus to test whether there is a QTL around the marker and for each r, the formula of LOD score for testing $H_0 : \mu_A = \mu_a$ $vs.$ $H_A : \mu_A \neq \mu_a$ is:

$$LOD(r) = -log_{10}\frac{L(\hat{\mu}_A = \hat{\mu}_a, \hat{\sigma}^2)}{L(\hat{\mu}_A, \hat{\mu}_a, \hat{\sigma}^2)}.$$

The second test is a special case of the first one. The LOD scores are then compared to a genome-wide threshold to infer the presence of a QTL. This method considers the fact that the LOD score is computed between markers. However, when using this method it is hard to combine the results for testing each marker and get a single estimation of the QTL location and effect.

Interval Mapping

Lander and Bostein (1989) introduced a significant improvement on QTL analysis by using the flanking markers to detect a QTL in experimental populations; this method is called "interval mapping". In this paper, they assume that there is no crossover interference for any pair of markers on the chromosomes under study and the phenotype is normally distributed. One has the information on the markers' locations and markers' genotypes. A backcross population is used here to explain this method. The phenotype of each individual follows a normal distribution with the mean equal to μ_A or μ_a depending on whether the QTL genotype is AA or Aa, and the variance σ^2 is defined as a common constant.

In a backcross population, there are two kinds of QTL genotypes and four possible genotypes at two flanking markers. Suppose the map distance between flanking markers is d. The map distance between the QTL and the left marker is d_L. According to Haldane's mapping function, the recombination rate between two markers is $r = \frac{1}{2}(1 - e^{-2d})$. The recombination rate between the QTL and the left marker is $r_L = \frac{1}{2}(1 - e^{-2d_L})$. And the recombination rate between the QTL and the right marker is $r_R = \frac{1}{2}(1 - e^{-2(d-d_L)}) = (r - r_L)(1 - 2r_L)$. We calculate the conditional probability for two possible genotypes of the QTL, given the flanking marker genotypes, by using a recombination rate. The results are shown in Table 1.1.

Suppose that for individuals with QTL genotypes AA, their phenotypes are distributed as $N(\mu_A, \sigma^2)$, and for individuals with QTL genotypes Aa, their phenotypes are distributed as $N(\mu_a, \sigma^2)$. Then for any given putative QTL location x, we can calculate the conditional probability assuming that the QTL genotype is AA for the i^{th} individual ($i = 1, \cdots, n$), given its flanking markers, and we note it as $P_i(x)$. Then the i^{th} individual's phenotype follows a mixture normal distribution:

Table 1.1: The conditional probability of a QTL genotype given two flanking makers' genotypes
is

marker genotype		QTL Genotypes	
left	right	Aa	AA
Aa	Aa	$(1-r_L)(1-r_R)/(1-r)$	$r_L r_R/(1-r)$
Aa	AA	$(1-r_L)r_R/r$	$r_L(1-r_R)/r$
AA	Aa	$r_L(1-r_R)/r$	$(1-r_L)r_R/(1-r)$
AA	AA	$r_L r_R)/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

$$y_i \sim P_i(x) \times N(\mu_A, \sigma^2) + (1 - P_i(x)) \times N(\mu_a, \sigma^2)$$

.

For each fixed location x, we use an EM algorithm to maximize the joint likelihood function and get the estimation for the unknown parameters (See Dempster *et al.* (1977)). Considering the null hypothesis that there is no single QTL on the chromosome, the LOD scores are calculated and plotted against x. The formula of LOD score for test $H_0 : \mu_A = \mu_a$ *vs.* $H_A : \mu_A \neq \mu_a$ is:

$$LOD = -log_{10}\frac{L(\hat{\mu}_A = \hat{\mu}_a, \hat{\sigma}^2)}{L(\hat{\mu}_A, \hat{\mu}_a, \hat{\sigma}^2)}.$$

We infer the presence of a QTL, if the LOD score at this position exceeds the genome-wide threshold.

The interval mapping method and the above MLE method are not identical. In the MLE method, we only consider the recombination rate between the QTL and one marker. But with interval mapping, we consider the distribution of QTL's genotype, given two flanking markers' information. The interval mapping method can give a precise estimate of the location and effect of a QTL. Therefore, many QTL statistical methods in the 1990s are the extensions based on the interval mapping methods. However, it has the drawback: it requires intensive computation compared to previous methods.

1.1.4 Likelihood-Based Multiple QTL Methods

Many traits may be influenced by multiple genes, so one QTL model is not sufficient to deal with this situation. We need to develop more complicated models because using one QTL model to detect multiple QTL data may fail to identify and estimate the multiple QTL locations. The detection power therefore is compromised and estimations of the QTL locations and their effects will be biased (Lander and Bostein (1989); Knapp (1991)). Sometimes "ghost QTL" may appear, in the sense that if there are two QTL on a chromosome evaluated with one QTL mapping method, you may detect a QTL located somewhere between two true QTL locations instead of detecting either one of them (Haley and Knott (1992); Martinez and Curnow (1992); Yi (2005)).

Multiple QTL can be mapped more accurately and more efficiently with a multiple QTL model. We will discuss three main likelihood-based methods for multiple QTL mappings: multiple linear regression, composite interval mapping (CIM) and multiple interval mapping (MIM).

Multiple Linear Regression Analysis

This multiple QTL method is an extension of the ANOVA method for one QTL model. In this method, the phenotypes of individuals are regressed on the markers' genotypes. The basic idea is that the effects of a QTL will be partially absorbed by linked markers (Stam (1991)). Cowen (1989) used stepwise selection and backward deletion techniques to select a class of markers, which are linked to a QTL. More recently, Doerge and Churchill (1996) described using forward selection and permutation tests to determine the number of markers in the model. However, when the distances between markers and QTL are large, only a small part of the QTL effect is absorbed by the markers. The power of detection thus becomes very small. We cannot estimate the precise locations and effects of a QTL using this method.

Composite interval mapping

With one QTL interval mapping, the likelihood function for a single QTL is assessed at each putative location on the chromosome. However, a QTL located somewhere else on the genome can have an interference effect. Jansen (1993), Zeng (1993) and Zeng (1994)

independently proposed the idea of combining interval mapping with multiple regression on markers' genotypes. Zeng (1994) named this method Composite Interval Mapping (CIM). The method is achieved by fitting one QTL interval mapping method and using part of the markers as co-factors to eliminate the effects of additional QTL. By fitting other genetic markers in the model as a control, it confines the test of one QTL to a region, which changes the problem from a multi-dimensional search to a one-dimensional search. Compared to one QTL interval mapping, CIM improves both the sensitivity and accuracy by including the markers, which may absorb the effects of other QTL. The parameters in the model are estimated by the expectation/conditional maximization (ECM) algorithm (Meng and Rubin (1993)).

The main challenge in this model lies in determining which markers to use as regressors. Jansen and Stam (1994) used backward deletion to pick up the subset of most significant markers with Akaike's Information Criterion (AIC). Zeng (1994) recommended that one include all the markers except those that are within 10 CM of the putative location.

Multiple interval mapping

If there are multiple QTL in the model, Lander and Botstein suggest detecting QTL one by one, i.e. they fix the position of the first QTL, then look for the next QTL location. This is a forward selection procedure (Miller (1990)). However, there is the drawback of a "ghost QTL" effect, in the sense that if there are two or more linked QTL, then interval mapping often gives a maximum LOD score at a location between the two QTL; see Haley and Knott (1992).

Kao *et al.* (1999) extend one QTL interval mapping model to a multiple QTL interval model. They use multiple marker intervals simultaneously to detect multiple putative QTL in the model. This method uses the general formulas derived by Kao and Zeng (1997) to obtain maximum likelihood estimates (MLEs) for the parameters. Compared with the regression method, this method gives accurate and precise locations and the effects of multiple QTL. However, the selection of the QTL involves multidimensional search, which is very computationally intensive.

1.1.5 Bayesian QTL Methods

Likelihood-based QTL methods detect the locations and effects of QTL mainly by maximizing the likelihood and evaluating the presence of QTL by using the LOD score. When computing confidence intervals, likelihood-based methods do not properly account for uncertainties in the parameters. With Bayesian methods, the prior information is incorporated into the analysis and the inferences are based on the marginal posterior distributions of the parameters, which are easy to interpret.

Satagopan *et al.* (1996) applied the standard Markov chain Monte Carlo (MCMC) method to map a given number of multiple QTL on the genome. MCMC is a Bayesian method commonly used to approximate a multi-dimensional integral of the likelihood function, which has no closed form. We generate a sequence of samples from the joint conditional probability distribution to get the integral. For the posterior probability of the parameter we are interested in, we sample each parameter from its conditional distribution given the rest of the parameters. The samples of the parameters are generated sequentially until the chains converge. Satagopan *et al.* (1996) used Gibbs sampling as well as Metropolis-Hastings algorithms to sample unknown parameters and missing data from their joint posterior distribution. The parameters were inferred based on their marginal posterior distributions, which can be obtained from the joint posterior distribution by integrating over the other unknowns. It is hard to get the exact integrations over multiple parameters, however, a Monte Carlo approximation is quite feasible for estimating the integrations. In the paper, the probability intervals for locations of multiple loci and their effects are discussed. This method accounts for the uncertainties in the parameters by considering the marginal posteriors, which average over such uncertainties in the parameters. The present paper also discusses the number of loci affecting the trait of interest. We estimate the number of QTL by fitting various models with different numbers of QTL, then we use a Bayes factor (Kass and Raftery (1995)) to compare these models.

The above Bayesian inference is complicated when the number of QTL is unknown. Essentially, the parameter space is the product of the spaces of different numbers of QTL. Most conventional techniques can't be applied. Green (1995) proposed using reversible jump

Markov chain Monte Carlo alogrithms specifically for such problems. This method combines the traditional MCMC algorithm with Metropolis-Hasting (Hastings (1970), Metropolis *et al.* (1953)) for jumping between different number of QTL. Satagopan and Yandell (1996) used a reversible jump MCMC to fit a multiple QTL model by including the number of QTL as unknown parameters. The locations, effects, and number of QTL can be estimated from the samples. Their application to the Brassica flowering data (Satagopan and Yandell (1996)) shows similar results compared to the results obtained using Bayes factors (Satagopan *et al.* (1996)). Because the reversible jump MCMC algorithm is very general and widely applicable, many effective approaches for detecting multiple (non-)epistatic QTL are based on it in different experimental populations or in pedigrees (Stephens and Fisch (1998); Sillanpaa and Arjas (1998); Thomas *et al.* (1997); Yi and S. (2000); Gaffney (2001); Yi and Xu (2002); Yi *et al.* (2003)). However this method also has its drawbacks: it is very poor to mix the chain for updating QTL locations and it is very slow for chain convergence.

The Bayesian approach above provides a sensible inferential framework for multiple QTL mapping. But it suffers from an intense computational burden. Berry (1998) has proposed another Bayesian model, which is also a Markov chain Monte Carlo method, but one with moderate computational speed. Usually the joint likelihood function is integrated over all other unknown parameters to get the posterior distributions of the number and the locations of QTL. When the number of parameters is large, the computation becomes very intensive. In Berry (1998), the ease of the computational burden was achieved by several approximations. Berry (1998) uses a first order approximation of the likelihood function, and the Laplace approximation to estimate the posterior distribution on the whole genome. The computation is improved through these approximations. Gibbs sampling is used to generate samples from the approximated posterior distributions. The number and the locations of QTL are inferred from the samples. The strength of this method is the moderate computation speed achieved by using fast approximations. However, this method is applied only in backcross populations and the accuracy of this approximation still requires further investigation.

Yi (2004) improved the efficiency of the reversible jump Markov chain Monte Carlo algorithm by using a unified Bayesian model selection framework for detecting multiple

QTL. This method is based on a composite space representation of the problem, which was developed by Carlin and Chib (1995). It provides a new viewpoint on the model selection problem. The advantage of this new method is that it allows Markov chain Monte Carlo simulation to be performed on a space of fixed dimension, thus avoiding the complexities of reversible jump technique. The Bayesian approach is finally simplified. The composite model space approach is extended to include epistatic effects in the model (Yi *et al.* (2005);Yi *et al.* (2007a)). They developed a computationally efficient Markov chain Monte Carlo algorithm using a Gibbs sampler and Metropolis-Hastings algorithm to study the posterior distribution of the parameters.

There are also some other Bayesian methods that can be used to calculate the posterior probabilities: numerical grid search integration, adaptive numerical integration, and importance sampling. Numerical grid search integration is a method that is used to approximate the integral function with no closed form by using a set of grid points. We obtain those grid points in a user-defined size domain for each nuisance parameter. We also know that the domains of the nuisance parameters for the integral likelihood is on the space Ω, and Ω can be arbitrarily large. If we truncate Ω to a reasonable rectangle size such that the likelihood would be very small outside of the rectangle, numerical grid search integration can divide this rectangle into many small cubes by the grid points we define, and get the integral value for each small cube. Then we can add up all the integral values on these small cubes to find the approximate value for the integral likelihood function. If the cubes are small enough, we should get a good approximation for the integration of the likelihood function. The disadvantage of this method is that it has a computationally intensive problem, so it is hard to apply it to high-throughput applications.

Adaptive numerical integration is a method used to approximate the integral over a multidimensional finite range by a recursive adaptive method, which divides the interval into two and compares the values given by Simpson's rule and the trapezium rule (Venables and Ripley (1994)). In R software, the "adapt" command in the package "adapt" is used to apply this method. It works well when models have only a few nuisance parameters. But when we apply this method to a model with many nuisance parameters or to a very complicated model, the adapt command in R software is very computationally intensive,

and can crash easily.

Importance sampling is a Monte Carlo method used to approximate an integral function by the average of ratios of likelihood density to the proposal density. A set of samples are evaluated to obtain the average ratios. We have to find a good proposal density, which is easy to simulate and "near" the density we want to integrate.

Sometimes, the traits we are interested in are binary responses. With most existing QTL mapping methods, the linkages between markers and QTL are tested with a simple chi-square test because of the binary traits. Xu (1996) proposed a composite interval mapping method, which treats a binary trait as the outcome from an underlying normally distributed liability. The quantitative liability is modelled by the usual QTL mapping method, since it is quantitative and continuous. Huang *et al.* (2007) have proposed a new Bayesian method which combines the unified Bayesian method and the liability model for studying binary traits. This Bayesian method uses all the markers on the entire genome simultaneously. Huang *et al.* (2007) developed the method for the case in which the QTL are located at the observed markers, and for the case in which QTL are located between markers. If the QTL are located between markers, the first method will lead to biased estimations. However, if the markers are dense enough, the first method will be quite accurate and could save much computation.

1.2 Microarrays Introduction

Microarray technology has become very popular in recent years and plays an increasingly important role in biomedical research. Disruptions or changes in genes can cause disease or morphological anomalies. By using microarray technology, we can detect changes in gene expression and prevent the genetic defects in advance. With microarray technology, we can measure thousands of genes simultaneously for different types of cells or tissues and use gene expression to describe their DNA information. Many statistical problems have arisen recently in the use of microarray data. Well-developed statistical methods that can assist us in locating the genes of interest are urgently needed.

A microarray data structure is shown in Figure 1.2 on the next page. In Figure 1.2, the top row denotes the names of the samples and the left column shows the names of the

genes. Each row in the gene expression matrix represents the expression values for each gene with respect to all samples. Each column in the gene expression matrix represents the expression values of each sample for all genes.

Spotted cDNA microarrays and oligonucleotide arrays (Affymetrix, Santa Clara CA) are two of the most commonly used gene expression arrays. In spotted cDNA microarray experiments, the ratio of red and green fluorescence intensity for each spot (gene) is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. The red (R, Cy5) labeled and green (G, Cy3) labeled mRNAs represent test and control samples, respectively. Probes are cDNA fragments attached on a solid support (a nylon or glass slide). The process works like this: first, the red and green labeled RNA samples are mixed and hybridized to the microarrays, which the supplier has spotted with cDNA from thousands of genes, each spot representing one gene. After hybridization, the red or green fluorescent signal from each spot is determined and the ratio of red to green is the primary measurement considered. If one gene has a signal closer to red, this means that gene is expressed at a higher level in the test sample than in the control sample. Newton *et al.* (2001), Dudoit *et al.* (2002) and Lee *et al.* (2000) represent some early representative papers for two-color microarrays.

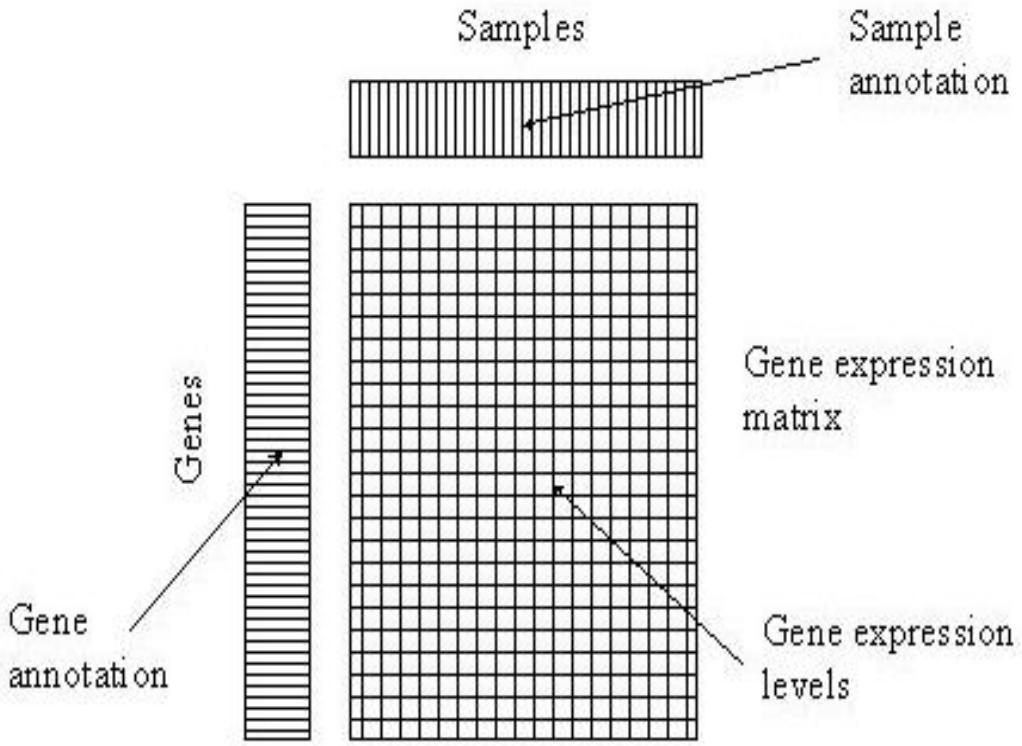

Figure 1.2: Microarray data structure.

In oligonucleotide arrays, instead of using one probe per gene, 11 to 20 probes are used to represent each gene (Lockhart *et al.* (1996)). Each probe represents a unique DNA fragment of one gene so a group of probes identifying a gene is called a probe set; in principle, one can obtain a better estimate of the expression level for the gene on probe-set arrays than for the gene on single-probe arrays. In Affymetrix technology, there is a perfect match (PM) probe for the target DNA sample, as well as a corresponding paired "mismatch" probe (MM). This mismatch probe contains only a single base change in the nucleotide located in the middle of the 25-base probe sequence; it is designed to measure non-specific hybridization as well as provide the information on background and cross-hybridization (Lipshutz *et al.* (1999)). A perfect match probe and its mismatch probe are called a probe-pair.

There are some differences between these two arrays. For spotted cDNA microarrays, one probe represents one gene; each array has two target samples or one target sample, and one reference sample; probe length on spotted cDNA microarrays varies. For oligonucleotide arrays, there are 11-20 probe-pairs per gene, each array has one target sample, and probe length is fixed with 25-mers (base pairs).

Statistical problems in this field involve microarray pre-processing like image analysis, background correction, expression quantification, normalization and quality assessment. There are also interesting problems when one is comparing two different conditions, like normal/disease, control/treatment, or when one is comparing more than two different conditions with microarray data. Statistical methods, like estimates and hypothesis testing, are applied to solve these problems. Exploratory analysis using microarrays is also important. Let us say we need to find a group of genes for a novel disease by using clustering or projection methods. There are also many other statistical problems and developed statistical methods in microarray analysis, which is not the focus in this dissertation and will not be discussed here.

1.3 eQTL Introudction

Quantitative geneticists are now interested in detecting expression quantitative trait loci (eQTL) for gene expression abundances because transcript abundances are considered to correlate with some important phenotypes. Transcript abundances can be treated as

a surrogate of phenotypes (Schadt *et al.* (2003)). eQTL methods have been developed to identify major-effect eQTL for transcripts by combining quantitative trait loci (QTL) mapping methods with microarray data, and "eQTL" are statistically significant peaks in a genome-wide scan for linkage analysis. In eQTL analysis, the experimental design is very similar to traditional quantitative trait loci analysis. The difference is that the expression values for the gene transcripts are treated as the phenotypes, so one must analyze thousands of phenotypes in eQTL analysis. Because of this difference, traditional QTL statistical methods, designed for testing (at most) tens of phenotypes, cannot be easily applied. Experimental design issues need to be addressed to handle these large data sets and new statistical methods are still being evaluated.

Brem *et al.* (2002) used the Wilcoxon-Mann-Whitney method for testing the significance of the linkage between each marker and transcript. This test has shown some promise in important biological situations and the resulting p-values for some transcripts are sufficiently small. Schadt *et al.* (2003) have used a traditional QTL interval mapping method for analyzing maize, mouse and human data sets. This likelihood-based approach can be used to obtain transcript-specific significance profile likelihood curves. However, those methods are still not well refined for problems like the potential increase in type I error by testing multiple markers, or power loss. Kendziorski *et al.* (2006) have proposed a Mixture-over-Markers (MOM) model to localize eQTL and have controlled the false discovery rate without sacrificing power. This method is a marker-based model. If the marker density is not sufficiently dense, the results for loci between markers may have some bias. New statistical methods are still needed to evaluate the eQTL data and optimize the test results.

Many eQTL studies based on the statistical methods mentioned above or some other very simple statistical methods have been published for many creatures, e.g. yeast (Brem *et al.* (2002); Yvert *et al.* (2003)), eucalyptus (Kirst *et al.* (2004)), mice (Schadt *et al.* (2003); Bystrykh *et al.* (2005); Chesler *et al.* (2005)), rats (Hubner *et al.* (2005)), maize (Schadt *et al.* (2003)) and humans (Morley *et al.* (2004); Monks *et al.* (2004); Hubner *et al.* (2005)). For those main regulated transcripts, results reported in several papers show that up to one-third of the significant genes are cis-acting, which means the gene expression values can be explained by the physical locations themselves. The rest of the significant genes are trans-

acting, which means that the gene expression values are regulated by other physical gene locations. Most of the cis-acting genes explain a greater proportion of expression variation than trans-acting genes. Trans-acting genes usually explain little variation individually, but we have more of them. This is similar to our expectation that DNA variation can affect a large portion of gene expression for that gene.

In summary, eQTL analysis has the potential to impact biological endeavors in a wide range of biomedical and agricultural fields. Applying traditional QTL methods to microarray data also gives us insight into gene networks, as well as their evolution. Because of the computational demands in eQTL analysis, the current statistical methods are not suitable for this high-through application. Well-developed and fast statistical methods are still needed to handle thousands of phenotypes efficiently.

1.4 Thesis Summary

In Chapter 1, we introduced QTL mapping and summarized some existing methods for detecting QTL. We also provided a brief introduction to microarray data and eQTL analysis.

In Chapter 2, Bayesian methods via the Laplace approximation for detecting single QTL are proposed. They can be easily applied to a backcross (BC) population and an F_2 intercross population. They can also be trivially extended to double haploid and other types of recombinant inbred strains. The applications to simulated data and real data demonstrate the high speed and high efficiency of these methods compared to alternative grid search integration, importance sampling, MCMC and adaptive quadrature methods. Our results also provide insight into the connection between the LOD curve and the posterior probability for linkage. In the application of our Bayesian method, we extend our approximate Bayesian linkage analysis approach to the expression quantitative trait loci (eQTL) model, in which microarray measurements of thousands of transcripts are examined for linkage to genomic regions. This approach uses the Laplace approximation to integrate over genetic model parameters (not including genomic position), and has been fully developed for different types of recombinant inbred crosses. The method is much faster than the more commonly-used Monte Carlo approaches, and thus is suitable for the extreme computational demands

of eQTL analysis. The approach offers highly interpretable direct posterior densities for linkage for each transcript at each genomic position. Biologically attractive priors involving explicit hyperparameters for probabilities of cis-acting and trans-acting QTL are easily incorporated.

In Chapter 3, we extend the one QTL Bayesian method via Laplace approximation method to the Bayesian method of detecting multiple QTL simultaneously. This joint multiple QTL Bayesian method has the advantage of providing the posterior probability at putative QTL locations and can detect QTL with interaction effects. The computation is intensive when you detect multiple QTL at the same time even using our efficient joint multiple QTL Bayesian method. Therefore, we also propose an iterative multiple QTL Bayesian method based on the Laplace approximation for detecting multiple QTL locations without the posterior probability calculation. The speed of this method is much faster than that of the joint multiple QTL Bayesian method. In this Chapter, we use the two QTL model as an example to demonstrate both our methods. We also apply our methods to simulation studies and real data analysis.

CHAPTER 2

One QTL Model

For the problem of mapping quantitative trait loci in experimental crosses, the interval mapping maximum likelihood approach of Lander and Bostein (1989) inspired a number of extensions, including regression approximations (Haley and Knott (1992)), composite interval mapping (CIM), multiple-QTL mapping (MQM)(Jansen (1993); Jansen and Stam (1994); Zeng (1993), Zeng (1994)) and multiple interval mapping (MIM) (Kao *et al.* (1999)).

The asymptotic results in Kong and Wright (1994) detailed non-standard behavior of maximum likelihood estimates for QTL positions. Moreoever, model selection remains a challenging and important aspect of linkage mapping, for which standard asymptotic approximations in traditional likelihood ratio testing may not work well. These are among the reasons for the popularity of Bayesian QTL mapping methods, which have an advantage in producing posterior densities for all model parameters.

However, most published Bayesian QTL approaches use Monte Carlo sampling of the parameter space (Satagopan *et al.* (1996); Berry (1998); Sillanpaa and Arjas (1998); Stephens and Fisch (1998); Yi and S. (2000), Yi and Xu (2001), Yi (2004), Huang *et al.* (2007)), which is too slow for high-throughput applications in which the analysis must be repeated thousands of times. The model introduced here formally applies to the single-QTL setting (per phenotype), and extensions to the multiple QTL setting are underway. Nonetheless, our model has an immediate application to the analysis of expression quantitative trait loci (eQTL), in which tens of thousands of transcripts are analyzed as phenotypes for linkage (Schadt *et al.* (2003)). Previous eQTL methods are based on likelihood (Kendziorski *et al.* (2006)), or are computationally intensive Bayesian approaches for which posteriors are evaluated at only a few hundred marker positions (Gelfond *et al.* (2007)). A computationally

efficient Bayesian eQTL approach would open up new avenues for research, enabling flexible incorporation of prior biological information.

The present chapter describes the mechanics of our approach in detail, which incorporates important simplifications in the model and in integration over nuisance parameters. Our method has utility beyond eQTL anaysis. For example, a fast Bayesian method can be used in sensitivity testing to various parameter settings. Other uses include empirical Bayesian methods in which the posterior linkage probabilities are used to evaluate the likelihood for population hyperparameters, for example in meta-analyses of multiple experimental crosses.

The major problem in linkage analysis concerns inference on the existence and position of a QTL. Accordingly, Bayesian QTL analysis fundamentally involves integration over nuisance parameters, i.e., any parameters other than the QTL position itself. As a Monte Carlo alternative to MCMC, we may consider importance sampling of the posterior of the likelihood in the vicinity of the maximum likelihood estimate (m.l.e.). Noting that there are relatively few nuisance parameters in experimental cross models, it also may be reasonable to consider direct numerical integration, including grid search integration (Thisted (Mar. 1998)) and adaptive quadrature (Venables and Ripley (1994)). However, as we demonstrate in Results below, none of these approaches is practical for high-throughput applications.

As a fundamentally different approach, we consider the *shape* of the likelihood in order to obtain insight into the problem. We note that the non-standard asymptotic behavior of the likelihood is confined to the QTL position estimate (Kong and Wright (1994)). At a fixed putative position, the likelihood for the nuisance parameters typically follows regularity conditions for standard inference (Azevedo-Filho and Shachter (1994)). As a consequence, integration over the nuisance parameters may employ the Laplace approximation, which essentially involves approximating the likelihood by an unscaled multivariate normal density (Crawford (1994)). Integration then becomes equivalent to determining the scaling factor, for which we will use the m.l.e. and analytic derivations of the Fisher information. Thus the necessary computation is of the same order as standard LOD approaches. The Laplace approximation has been used to speed up an evaluation step in the MCMC method of Berry (1998), but otherwise has been largely overlooked in this setting.

Here we employ the Laplace approximation to obtain the linkage posterior for backcross (BC) and F2 intercross data. The backcross approach also applies to double haploid populations, and essentially applies to recombinant inbred data sets, albeit with a higher effective recombination rate (Haldane and Waddington (1931)). For completeness, we provide comparisons MCMC and alternative integration approaches as described above, demonstrating that the Laplace approximation is highly accurate and, due to its speed, uniquely suitable for high-throughput applications. Simulations indicate that the advantages hold over a wide range of sample sizes, heritability, and other conditions. We further illustrate our approach by analyzing a real F2 mouse data set for plasma HDL cholesterol concentration (HDL) (Ishimori *et al.* (2004)) and use the budding yeast data from Brem *et al.* (2002) to illustrate the eQTL analysis.

2.1 Methods

Throughout this dissertation we use the normal linear phenotype model commonly applied to quantitative trait data (Lander and Bostein (1989)). However, the general approach is applicable to a wide variety of parametric phenotype models, and the vast majority of QTL models fall within the exponential family (Wright and Kong (1997)).

Let y_i denote the phenotype for the i^{th} individual. For a BC individual, we have the model

$$y_i = \mu + a \cdot g_i(x^*) + \epsilon_i, \tag{2.1}$$

where a is the additive effect of the QTL; $g_i(x)$ is a numerical representation of the genotype for the i^{th} individual at position x, x^* is the true QTL location, and ϵ_i is residual error, distributed $N(0, \sigma^2)$. We code $g_i(x)$ as 1 or -1 according to whether the genotype at x is AA (homozygote) or Aa (heterozygote). We use $\boldsymbol{\beta} = \{\mu, a, \sigma^2\}$ to represent the nuisance parameters, occupying a possibly finite region Ω for which the prior $p(\boldsymbol{\beta}) > 0$. We wish to obtain the posterior probability of the QTL at any gene location x, given phenotypes and marker genotype data,

$$p(x|data) = \frac{p(x)p(data|x)}{p(data)} = \frac{p(x)\int_\Omega p(data,\boldsymbol{\beta}|x)d\boldsymbol{\beta}}{p(data)} = \frac{p(x)\int_\Omega p(\boldsymbol{\beta})p(data|x,\boldsymbol{\beta})d\boldsymbol{\beta}}{p(data)}. \quad (2.2)$$

Here x denotes the true QTL position, so for example the location prior $p(x)$ will be understood to mean $p(x^* = x)$. This prior is intentionally flexible, as for future applications it might be sensible to consider prior information from previous studies, or to place mass only on the genomic positions of genes, implicitly favoring gene-rich genomic regions. Our goal is to enable direct probability statements for the posterior of x at each position, so that the posterior for entire regions/chromosomes may be obtained via summation or integration. In contrast, numerous Bayesian QTL methods are inherently dependent on Bayes Factors for inference (Satagopan *et al.* (1996); Berry (1998) etc.), for which evaluation of the evidence is less formal (Kass and Raftery (1995)). Nonetheless, Bayes Factors may also be easily obtained from our approach (see Discussion).

The right-hand side of (2.2) follows from the assumption of independence of QTL position and effect size, $p(x,\boldsymbol{\beta}) = p(x)\ p(\boldsymbol{\beta})$. We will denote the marker positions by the vector $\mathbf{x}_\mathrm{m}$, and the markers flanking x by $\{x_{left}, x_{right}\}$. The quantity $p(data|x,\boldsymbol{\beta})$ is the ordinary interval mapping likelihood for n individuals:

$$\begin{aligned} p(data|x,\boldsymbol{\beta}) &= p(g(\mathrm{x_m}))\prod_{i=1}^{n}\Big[\sum_{k=-1,1} p\big(y_i|g_i(x)=k,x,\boldsymbol{\beta},g_i(x_{left}),g_i(x_{right})\big) \\ &\times\ p\big(g_i(x)=k|\boldsymbol{\beta},g_i(x_{left}),g_i(x_{right})\big)\Big], \end{aligned} \quad (2.3)$$

for which we use model (2.2) and Haldane's map function for genotype probabilities.

Thus far, our presentation is simply a standard Bayesian outline of the problem. In contrast to other Bayesian QTL approaches (e.g. Satagopan *et al.* (1996); Berry (1998); Sillanpaa and Arjas (1998); Stephens and Fisch (1998); Yi and S. (2000), Yi and Xu (2001), Yi (2004), Huang *et al.* (2007)), however, we state the null hypothesis in terms of the QTL position x^*. If x^* is on the chromosome or chromosomes under study, the alternative hypothesis holds. Otherwise, the null hypothesis holds, which we denote H_0: $x^* = \infty$ (Doerge *et al.* (1997)). The more commonly-used form of null hypothesis, dating at least to

Lander and Bostein (1989), is a *no-gene* null specified in terms of the nuisance parameters as $\boldsymbol{\beta} \in \Omega_0 \subset \Omega$. The latter approach enables pointwise significance testing to follow standard likelihood ratio approximations in nested models (Lander and Bostein (1989)). However, in a Bayesian setting there is no inherent reason to favor this specification. Note that our null hypothesis can accommodate the situation where no gene exists - as the sample size increases, evidence will accrue that the effect size is negligibly small. In practical terms, for likelihood ratio testing the two forms of null hypotheses may be very similar, as the maximum null likelihood typically represents a similar fit to the data using either form. An exception occurs when a QTL with large effect is present on a chromosome other than the one under study, causing a bimodal phenotype distribution. Indeed, this situation is examined in Lander and Bostein (1989) as an example where no-gene specification produces poor inference. However, this situation presents no conceptual difficulty for our approach, because the possibility is explicitly considered that an unobserved QTL may produce such a phenotype mixture distribution.

A second and important advantage to our null hypothesis specification is that inference for x will be relatively insensitive to the prior for $\boldsymbol{\beta}$, because $p(\boldsymbol{\beta})$ appears in both null and alternative terms in $p(data)$. In contrast, when using the no-gene null hypothesis, inference can be highly sensitive to the prior for $\boldsymbol{\beta}$, where the subspace Ω_0 is typically of lower dimension than Ω. We use a flat (proper) prior in our illustrations of the Bayesian approach, $p(\boldsymbol{\beta}) = \frac{1}{|\Omega|}$. Thus Ω must technically be finite. However, for realistic sample sizes, we can let Ω get arbitrarily large, with essentially no change in our inference. This phenomenon is illustrated in the Simulations section.

Using the assumed prior for $\boldsymbol{\beta}$, the integral in the numerator of (2.2) becomes

$$\int_{\Omega} p(\boldsymbol{\beta}) p(data|x, \boldsymbol{\beta}) d\boldsymbol{\beta} = \frac{1}{|\Omega|} \int_{\Omega} p(data|x, \boldsymbol{\beta}) d\boldsymbol{\beta} = \frac{1}{|\Omega|} C(x), \tag{2.4}$$

where $C(x)$ is the integrated likelihood for a fixed x. The denominator of (2.2) is

$$p(data) = \int_{x'} p(x') \{ \int_{\Omega} p(\boldsymbol{\beta}) p(data|x, \boldsymbol{\beta}) d\boldsymbol{\beta} \} dx' = \frac{1}{|\Omega|} \int_{x'} p(x') C(x') dx', \tag{2.5}$$

so the $\frac{1}{|\Omega|}$ term cancels out in numerator and denominator. We obtain

$$p(x|data) = \frac{p(x)C(x)}{\int_{x'} p(x')C(x')dx'} = \frac{p(x)C(x)}{\int_{x'<\infty} p(x')C(x')dx' + p(\infty)C(\infty)}, \tag{2.6}$$

where the denominator is partitioned into H_A and H_0 portions, and $p(\infty)$ is the prior for $H_0 : \ x^* = \infty$.

For notational simplicity, we use integral notation for summing over x positions. In practice, the prior on x may be either continuous or discrete. Note that (2.6) neatly decomposes the posterior into $p(x)$ and $C(x)$ terms. Thus if the prior $p(x)$ is changed or updated from external sources, the posterior may be easily computed with no need to recompute $C(x)$. In this dissertation, we will use a discrete uniform $p(x)$ over a grid with respect to genetic map position. However, this choice of prior entails no loss of generality.

Finally, it simply remains to obtain $C(x)$ for each x, including the null value $C(\infty)$. No analytic solution is available, and we will use the results of a numerical grid search as the gold standard, to which we compare our proposed Laplace approximation, as well as a crude version of the Laplace approximation that is even more computationally efficient. For completeness, we also examine alternate methods for evaluating the integral, including adaptive numerical quadrature, importance sampling, and Markov chain Monte Carlo (MCMC) sampling.

2.1.1 The Laplace Approximation

We focus on a single chromosome, with $H_0 : x^* = \infty$ corresponding to the hypothesis that the QTL is unlinked to the chromosome (although the approach is just as easily applied to an entire genome scan). For fixed x, we define $f(\boldsymbol{\beta}) = p(data|x, \boldsymbol{\beta})$. The applicability of the Laplace approximation relies on standard behavior for the log-likelihood for large sample sizes: the function is continuous, unimodal, twice differentiable, with a maximum in the interior of Ω (Azevedo-Filho and Shachter (1994)). The Laplace approximation may be motivated by a Taylor expansion at $\hat{\boldsymbol{\beta}}$ for a fixed x:

$$log(f(\boldsymbol{\beta})) = log(f(\hat{\boldsymbol{\beta}})) - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\Sigma}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + O(||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^3). \tag{2.7}$$

The m.l.e. $\hat{\boldsymbol{\beta}}$ may be obtained using a standard maximization routine such as E-M, as is routinely performed in standard interval mapping. $\hat{\Sigma} = I^{-1}(\hat{\boldsymbol{\beta}})$ is obtained by inverting the analytically-derived information matrix at $\hat{\boldsymbol{\beta}}$.

After exponentiating both sides and integrating over $\boldsymbol{\beta}$, we obtain

$$C(x) = \int_{\beta \in \Omega} f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \int f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx f(\hat{\boldsymbol{\beta}})(2\pi)^{dim(\boldsymbol{\beta})/2} |\hat{\Sigma}|^{1/2} \equiv \hat{C}(x), \tag{2.8}$$

where the indefinite integral assumes the space Ω is "large," and the $(2\pi)^{dim(\boldsymbol{\beta})/2}|\hat{\Sigma}|^{1/2}$ term arises from integration over a multivariate normal density with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\Sigma}$. Finally, we substitute $\hat{C}(x)$ for $C(x)$ in equation (2.6) for $x < \infty$.

Our Laplace approximation is already very fast, but can be made even faster with a slight decrease in accuracy if the posterior tends to concentrate in a small genomic region. Under this scenario, we may replace $|\hat{\Sigma}(x)|^{1/2}$ by the single estimate $|\hat{\Sigma}|^{1/2}$ evaluated at the maximum posterior x location, because the uncertainty in $\boldsymbol{\beta}$ does not vary much in that region. We refer to this approach as our *Laplace fixed* approximation

2.1.2 Approximating the null integrated likelihood

For the null value $C(\infty)$, the Laplace method may technically be applied, and will be asymptotically accurate. However, we have performed simulations demonstrating that the full Laplace approximation does not perform well under the null hypothesis for realistic sample sizes when a true gene exists, but outside of the genomic region under study. Note that this is not a problem for the Bayesian approach, but affects the accuracy of the Laplace approximation. The difficulty is illustrated in Figure 2.1, which shows likelihood contours for $\{a, \sigma^2\}$ for two choices of μ when n=100. The null likelihood is a mixture of two normal densities, with each of the two genotype probabilities in equation (2.3) replaced by 1/2. In addition to the curvature in the likelihood contours, the likelihood can remain relatively high and flat spanning $a = 0$, and it is difficult to prescribe a parameter transformation that will make the likelihood approximately normal in shape. Furthermore, if such a transformation was available, it would be non-linear, and difficult to transform back to integration over the original Ω.

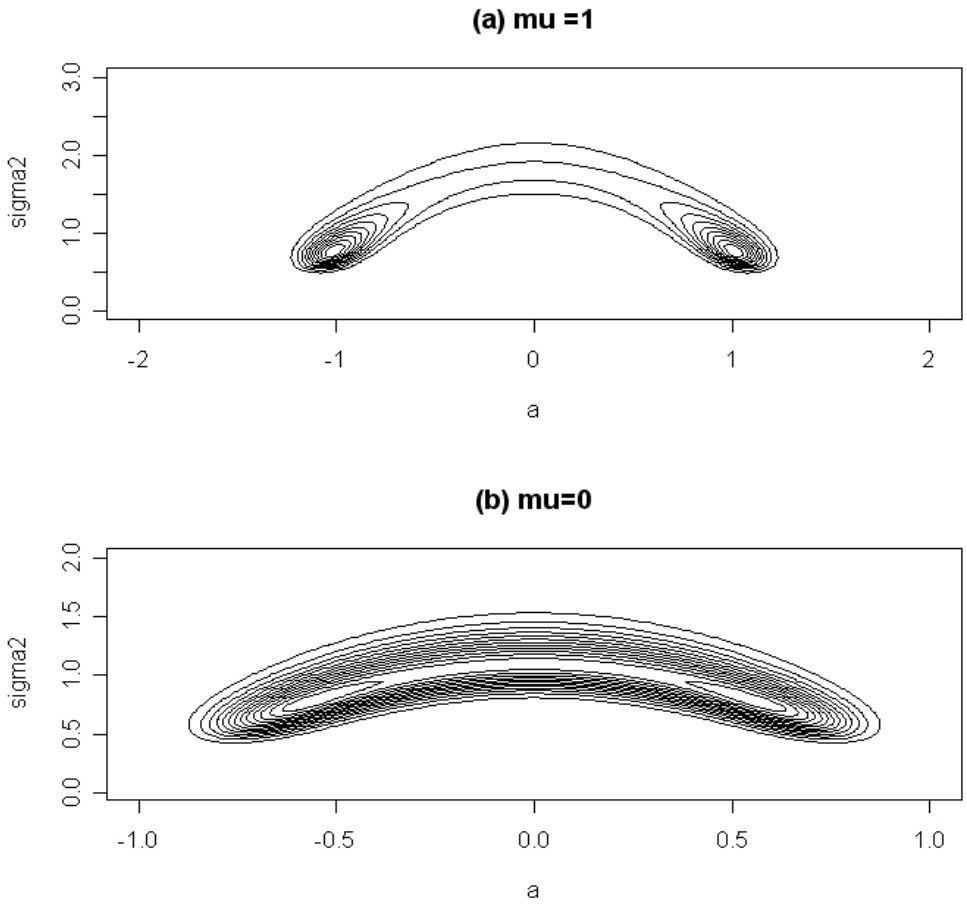

Figure 2.1: Contour plot of Laplace approximation for BC data under the null hypothesis: (a) n (sample size)=100, $\mu = 1$ (b) n (sample size)=100, $\mu = 0$.

One approach to this problem would be to apply numerical integration over Ω. However, we have devised the following approximation requiring integration over only one parameter, using the fact that the Laplace approximation for $\{\mu, \sigma^2\}$ works well for a fixed a (see Figure 2.2 (b) for an illustration). Define $f(a, \mu, \sigma^2) = p(data|x = \infty, a, \mu, \sigma^2)$, and $\hat{\mu}_a, \hat{\sigma_a^2}$ (obtained numerically) as the conditional m.l.e.s for fixed a, with corresponding covariance matrix estimate $\hat{\Sigma}_a$ on the restricted space. We then have the *improved null* Laplace approximation

$$\hat{C}(\infty) = \int_a [\int\int_{\mu,\sigma^2} f(a; \mu, \sigma^2) d\mu d\sigma^2] da \tag{2.9}$$

$$= \int_a f(a, \hat{\mu}_a, \hat{\sigma_a^2}) 2\pi |\hat{\Sigma}_a|^{1/2} da. \tag{2.10}$$

This *improved null* can be sped up with a further approximation. For fixed a, model (2.1) implies $E(Y|a) = \mu$ and $\sigma^2 = \text{var}(Y|a) - a^2$. For small to moderate a, the values y are approximately normal, with approximate conditional m.l.e.s $\hat{\mu}_a = \bar{y}$, $\hat{\sigma}_a^2 = s_y^2(n-1)/n - a^2/4$. The variance matrix terms are $\widehat{\text{var}}(\hat{\mu}_a) = s_y^2/n$, $\widehat{\text{var}}(\hat{\sigma}_a^2) = 2s_y^4/(n-1)$ and covariances=0. For larger a, we find empirically that the m.l.e. approximations continue to work well, and the covariance of the sample mean and variance remains zero (because the distribution of y is symmetric). These further approximations are used in equation (2.9) and termed the *fast null* Laplace approximation.

The accuracy of the all three Laplace null approximations is compared to that of a numerical grid search in Figure 2.2 for 20 simulations under the model $\{\mu, a, \sigma^2\} = \{0, 0, 1\}$ for $n = 100$. The *naive null* Laplace performs poorly (in Figure 2.2 (a)), and in fact will often not compute at all due to numerical instability arising from flat regions in the likelihood. In contrast, the *improved null* and *fast null* approximations are quite accurate (in Figure 2.2 (b) and (c)). The simulated results of all three Laplace approximations under other nuisance parameter values, such as $\{\mu, a, \sigma^2\} = \{0, 0.5, 1\}$ for $n = 100$, are similar and shown in Figure 2.3.

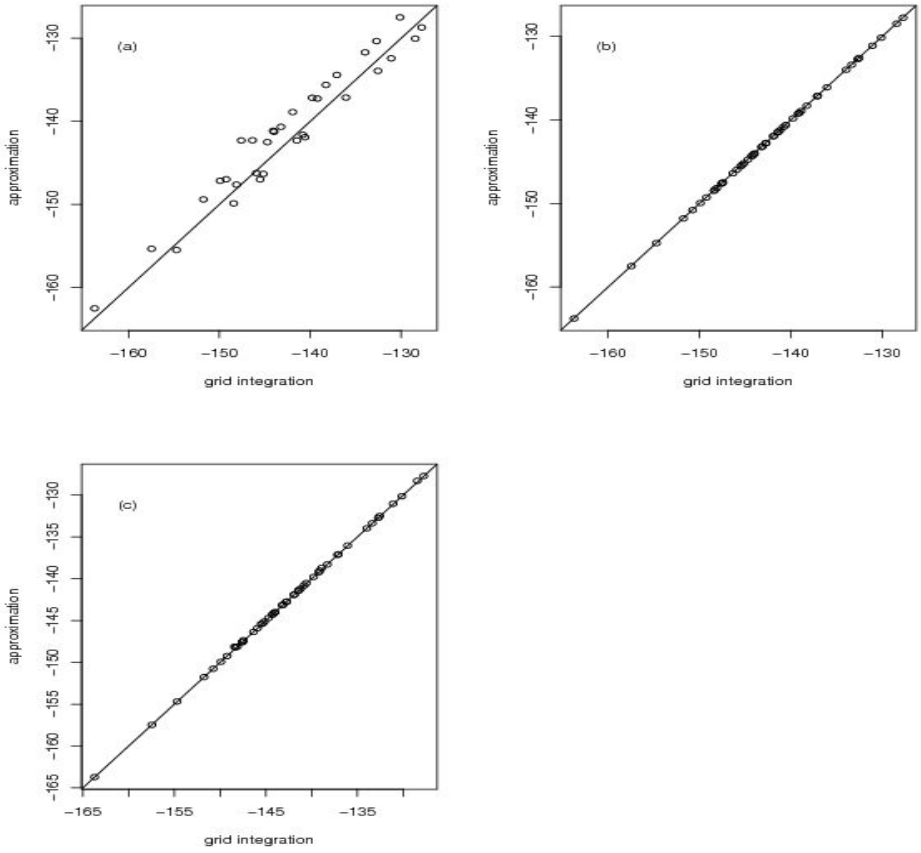

Figure 2.2: BC data simulation results for $\{\mu, a, \sigma^2\} = \{0, 0, 1\}$ and $n = 100$ under the null hypothesis: (a) *naive null* Laplace approximation (b) *improved null* Laplace approximation (c) *fast null* Laplace approximation.

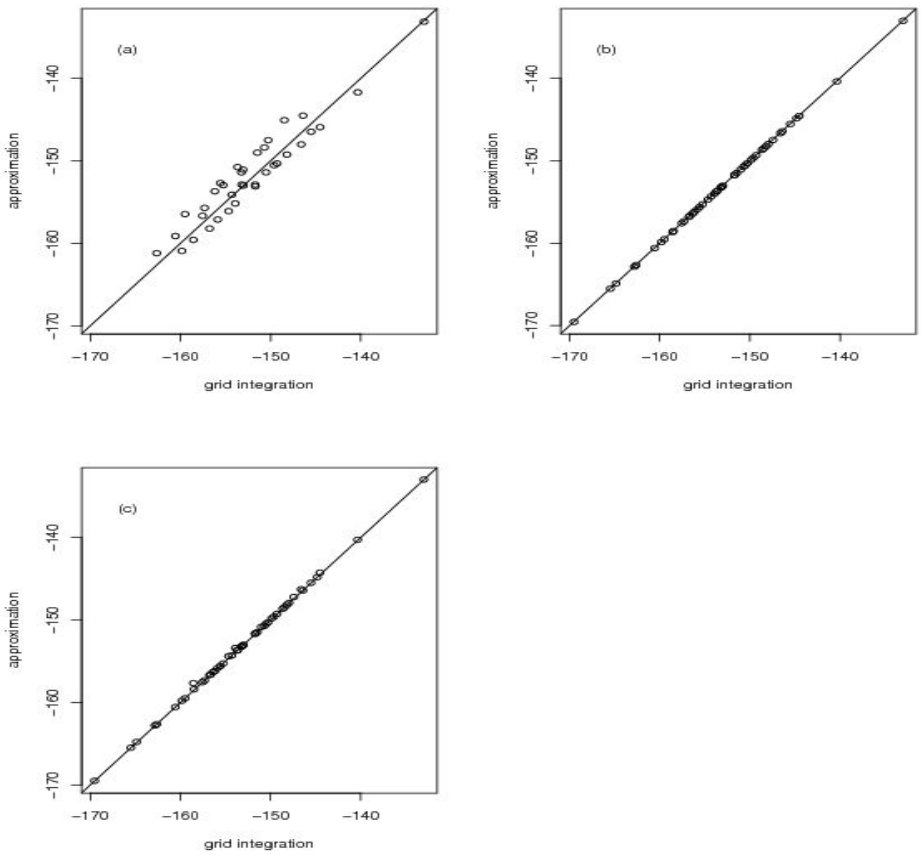

Figure 2.3: BC data simulation results for $\{\mu, a, \sigma^2\} = \{0, 0.5, 1\}$ and $n = 100$ under the null hypothesis: (a) *naive null* Laplace approximation (b) *improved null* Laplace approximation (c) *fast null* Laplace approximation.

2.1.3 Extension to F2 populations

The Laplace approximation for F2 populations is somewhat more complicated, but straightforward. The corresponding phenotype model is:

$$y_i = \mu + a \cdot g_i(x^*) + d \cdot (1 - |g_i(x^*)|) + \epsilon_i, \tag{2.11}$$

where a and d are the additive and dominance effects for the QTL; $g_i(x) = 1$ for genotype AA at x, 0 for genotype Aa, and -1 for genotype aa; and x^* is the true QTL position. The likelihood follows the same form as (2.3), except that the summation is now over genotypes -1, 0, and 1.

We use the Laplace approximation to estimate $C(x)$ under the alternative hypothesis and $C(\infty)$ under the null hypothesis. Because the alternative likelihood function for F2 is unimodal, we can obtain the accurate estimation for $C(x)$, which is similar to the estimation of $C(x)$ for BC population below:

$$C(x) = \int_{\beta \in \Omega} f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \int f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx f(\hat{\boldsymbol{\beta}})(2\pi)^{dim(\boldsymbol{\beta})/2} |\hat{\Sigma}|^{1/2} \equiv \hat{C}(x), \tag{2.12}$$

For estimation of $C(\infty)$, the accuracy of the *naive null* Laplace and *fast null* Laplace approximations is compared to that of a numerical grid search in Figure 2.4 for 20 simulations. Under the model $\{\mu, a, d, \sigma^2\} = \{0, 0, 0, 1\}$ for $n = 100$, the *naive null* Laplace again performs poorly and will not compute approximately 40% of the time due to numerical instability (in Figure 2.4 (a)). In contrast, the *fast null* Laplace approximation is quite accurate (in Figure 2.4 (b)). We also simulate 20 data sets under the model $\{\mu, a, d, \sigma^2\} = \{0, 0.5, 0.5, 1\}$ when $n = 100$ for the accuracy of the *naive null* Laplace and *fast null* Laplace approximations respectively and obtain similar results in Figure 2.4 (c), (d).

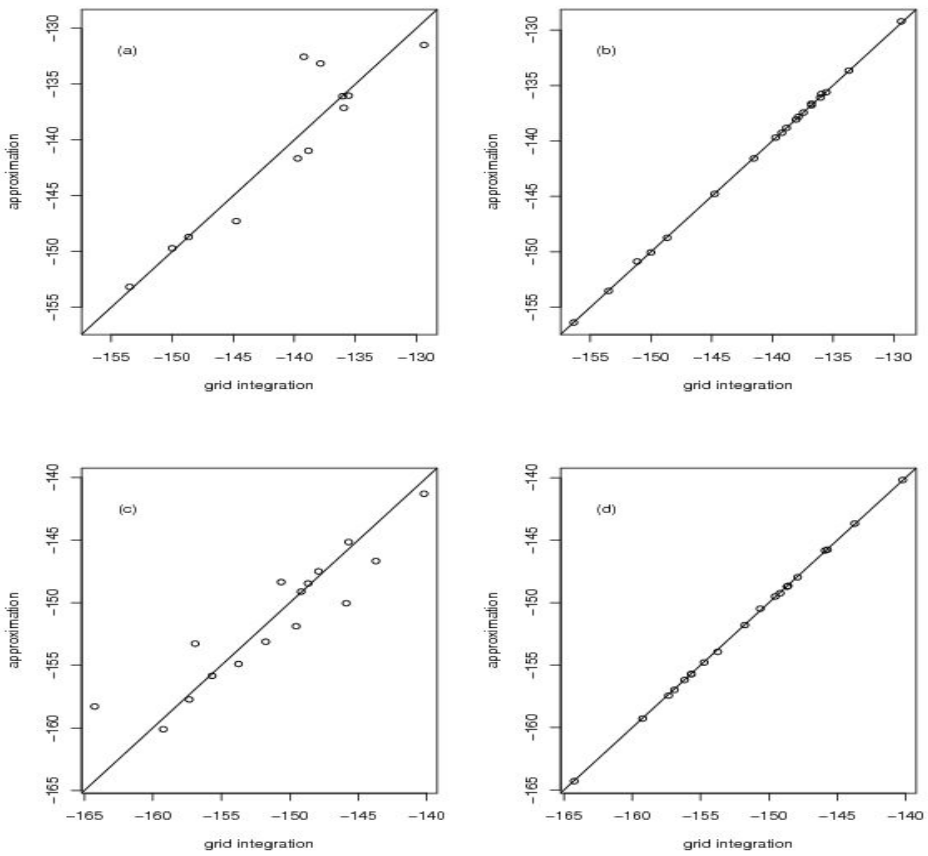

Figure 2.4: F2 data simulation results for $n = 100$ under the null hypothesis: (a) *naive null* Laplace approximation (b)*fast null* Laplace approximation under $\{\mu, a, d, \sigma^2\} = \{0, 0, 0, 1\}$; (c) *naive null* Laplace approximation (d)*fast null* Laplace approximation under $\{\mu, a, d, \sigma^2\} = \{0, 0.5, 0.5, 1\}$.

2.2 Relationship between Linkage Posterior Probability and LOD Score

In traditional QTL mapping, LOD score is usually the test statistic used to detect QTL locations. Posterior probability has an advantage when interpreting results for estimating QTL positions, as well as the advantages mentioned in the Introduction, so Bayesian methods are enjoying greater popularity in current QTL mapping. We are interested in the relationship between LOD scores and linkage posterior probability. Once we know the relationship between these two statistics, we can easily get the linkage posterior probability for studies which use LOD scores to detect QTL locations. Our proposed Bayesian method via the Laplace approximation can be easily applied and affords insight into the relationship using the same hypothesis as the usual QTL mapping methods use (Lander and Bostein (1989)).

By using our Bayesian method and the "no-gene" null hypotheses that $H_0 : a = 0$ for BC and $H_0 : a = d = 0$ for F2, we want to find the relationship between the LOD score and our linkage posterior probability. The priors of the nuisance parameters $\boldsymbol{\beta}$ are assumed to be improperly uniform distributed, that is, their domains are from $-\infty$ to ∞, so we know that $p(\beta) = 1$ ($\boldsymbol{\beta} = \{\mu, a, \sigma^2\}$ for BC; $\boldsymbol{\beta} = \{\mu, a, d, \sigma^2\}$ for F2). The following is the derivation to find the relationship:

$$p(x|data) = \frac{p(x)p(data|x)}{p(data)} = \frac{p(x)\int_{-\infty}^{\infty} p(data, \boldsymbol{\beta}|x)d\boldsymbol{\beta}}{p(data)} = \frac{p(x)\int_{-\infty}^{\infty} p(\boldsymbol{\beta})p(data|x, \boldsymbol{\beta})d\boldsymbol{\beta}}{p(data)}.$$

The integration part in the numerator is

$$\int_{-\infty}^{\infty} p(\boldsymbol{\beta})p(data|x, \boldsymbol{\beta})d\boldsymbol{\beta} = \int_{-\infty}^{\infty} p(data|x, \boldsymbol{\beta})d\boldsymbol{\beta} = C(x). \tag{2.13}$$

And its denominator is

$$p(data) = \int_{x'} p(x')\{\int_{-\infty}^{\infty} p(\boldsymbol{\beta})p(data|x', \boldsymbol{\beta})d\boldsymbol{\beta}\}dx' = \int_{x'} p(x')C(x')dx'. \tag{2.14}$$

Finally we have

$$p(x|data) = \frac{p(x)C(x)}{\int_{x'} p(x')C(x')dx'} = \frac{p(x)C(x)}{\int_{x'<\infty} p(x')C(x')dx' + p(H_0)C(H_0)}. \tag{2.15}$$

For fixed x, we define $f(\boldsymbol{\beta}) = p(data|x, \boldsymbol{\beta})$. By using the Laplace approximation at the MLE of the nuisance parameters, we get

$$C(x) = \int_{\beta\in(-\infty,\infty)} f(\boldsymbol{\beta})d\boldsymbol{\beta} \approx f(\hat{\boldsymbol{\beta}})(2\pi)^{dim(\boldsymbol{\beta})/2}|\hat{\Sigma}|^{1/2}. \equiv \hat{C}(x). \tag{2.16}$$

.

Similarly, we can obtain the estimate of $C(H_0)$. Then the posterior probability of the putative QTL location given data in this method is:

$$p(x|data) \approx \frac{p(x)f(\hat{\beta})(2\pi)^{dim(\beta)/2}|\hat{\Sigma}|^{1/2}}{\sum_{x'<\infty} p(x')f(\hat{\beta})(2\pi)^{dim(\beta)/2}|\hat{\Sigma})|^{1/2} + p(H_0)f(\hat{\beta}_0|H_0)(2\pi)^{dim(\beta_0)/2}|\hat{\Sigma}(H_0)|^{1/2}},$$

where $\beta_0 = \{\mu, \sigma^2\}$; $\hat{\beta}_0$ is the MLE of β_0.

Because LOD(x) $= log_{10}\frac{f(\hat{\beta})}{f(\hat{\beta}_0|H_0)}$, the above equation can be rewritten as:

$$\begin{aligned} p(x|data) &\approx \frac{p(x)f(\hat{\beta})(2\pi)^{dim(\beta)/2}|\hat{\Sigma}|^{1/2}}{\sum_{x'<\infty} p(x')f(\hat{\beta})(2\pi)^{dim(\beta)/2}|\hat{\Sigma})|^{1/2} + p(H_0)f(\hat{\beta}_0|H_0)(2\pi)^{dim(\beta_0)/2}|\hat{\Sigma}(H_0)|^{1/2}} \\ &= \frac{p(x)10^{LOD(x)}(2\pi)^{dim(\beta)/2}|\hat{\Sigma}|^{1/2}}{\sum_{x'<\infty} p(x')10^{LOD(x')}(2\pi)^{dim(\beta)/2}|\hat{\Sigma})|^{1/2} + p(H_0)(2\pi)^{dim(\beta_0)/2}|\hat{\Sigma}(H_0)|^{1/2}}. \end{aligned}$$

We know the linkage posterior probability $p(H_A|data) = \sum_{x<\infty} p(x|data)$. By using our method, we can easily get the posterior probability for any putative QTL location given

data and also provide insight into the connection between the LOD curve and the posterior probability for linkage.

2.3 Simulation Studies

We conducted the simulation studies for BC and F2 intercross populations, respectively, to evaluate the performance of our proposed methods and other existing methods. For the numerical grid search method, the domains we chose for nuisance parameters μ, a, d are from $-20/7$ to $20/7$ with grid size $1/7$. As for σ^2, the domain is from 0 to $20/7$ with grid size $1/7$. For the importance sampling method, the proposal densities we chose for μ, a and d are normal distributions using the MLEs of the parameters and standard errors. The proposal density of σ^2 is inverse gamma. We then use the expected ratios of likelihood to the proposal densities to obtain the numerical integral estimate. For the sampling process in the MCMC method, we burn in the first 10,000 sweeps of the chain and then we perform an additional 100,000 MCMC sweeps. The final samples were selected every 100 sweeps to reduce the correlation, resulting in 1000 samples from the posterior distribution. The approximated posterior distribution is calculated based on these samples. In R software, the "adapt" command in the package "adapt" is used for computing the adaptive quadrature method. All the simulation results below except the adapt quadrature method result are running in a Linux PC with cpu: Xeon 2.8 GHz by using a C program.

2.3.1 BC QTL Data

Our simulation study for BC is based on a 100 cM chromosome, with only one QTL for each individual. For data generation, we simulate marker genotypes and QTL genotype given the QTL location of all individuals by using the first order Markov chain (a standard assumption of genetics analysis), and the transition probabilities are the recombination rates between two markers or between one marker and the QTL. The phenotype of each individual was generated from a normal distribution. The mean value of the normal distribution depends only on the QTL genotype of that individual and the standard deviation is a fixed number we specify. After we have the simulated data, we evaluate the posterior probabilities at 100 possible QTL locations x, which are generated uniformly on the chromosome under

study. Six methods are used (see Statistical Methods) to get the posterior probabilities for those possible positions. Since the numerical grid search method is treated as the gold standard, we compare its result to the results of the remaining five methods and calculate the errors of those five methods respectively by using the following error formula:

$$\frac{\sum_{all x_i} |\widehat{p}(x_i|data) - p(x_i|data)| + |\widehat{p}(\infty|data) - p(\infty|data)|}{2}, \tag{2.17}$$

where $\hat{p}(.)$ means the posterior probabilities from the method we intend to evaluate and $p(.)$ means the posterior probabilities of the gold standard numerical grid search integration method. We simulate 100 different data sets to get the average error. We also record the average speed of each method by running 100 different simulated data sets.

In order to test the speed and the error under a wider variety of conditions including: the true location of the QTL, the number of individuals, the number of markers on the chromosome, and the heritability effect. We evaluate speed and error under 64 conditions of the following combinations: QTL location = 10.1cM, 45.1cM; the number of individuals equals to 100 or 200; the number of equally dense markers is 5, 10, 20, or 100; the heritability effect is 0, 0.05, 0.1 or 0.15 (therefore, $\boldsymbol{\beta} = \{0, a, 1\}$, where $a = 0$, 0.2294, 0.3333 or 0.4201, which accords with heridity effect 0, 0.05, 0.1 or 0.15 respectively.). We generate 100 different simulated data sets for each of 64 conditions to get more accurate average error and speed. We choose 0.9 as the prior for H_0 (no QTL on the chromosome under study). The prior for the QTL at each location is specified as $\frac{0.1}{100}$.

2.3.2 F2 QTL Data

Our simulation study for F2 is similar to the simulation process for BC. To study the effect of our proposed methods, we test the accuracy and the speed for these methods under 64 different conditions, which are the same as the conditions in BC simulation. Therefore, $\boldsymbol{\beta} = \{0, a, 0, 1\}$, where $a = 0$, 0.3244, 0.4714 or 0.5941, which accords with heritability effect 0, 0.05, 0.1 or 0.15, respectively. The priors for H_0 and possible QTL locations are defined as the same as the priors in the BC data simulation.

Because the speed of the grid search method in F2 is very slow, we only simulate 10

different data sets to get the average errors for those five methods under each condition. We also record the average speed of the grid search method by running 10 different simulated data sets for each condition. For the remaining five methods, we run 100 simulated data sets to get the average speeds and errors under each condition.

2.3.3 Simulation Results

Speed

We have mentioned that multiple data sets are simulated to get the average speed for each method under each condition. The results are displayed in Table 1 to Table 4 for BC data sets and in Table 5 to Table 8 for F2 data sets. Table 1 shows the average speeds for simulated BC data sets with sample size 100 and QTL location 10.1 cM. We found that the grid search method with average speed around 60 seconds is much slower than all other methods. *Laplace fixed* approximation is the method with the highest speed. The speed of the Laplace approximation method is very close to the speed of *Laplace fixed* approximation method. Both are less than 0.1 seconds, and are much faster than the speed of the grid search method. The importance sampling and MCMC methods have moderate speeds compared with the others. We also conclude that the speed is not affected by the heritability or the inter-marker distance (or the number of markers) on the chromosome. Tables 2, 3, and 4 display the speed results for different choices of the sample size and the QTL location. We compared Tables 1, 2 with Tables 3, 4 and found that the sample size greatly affects the speed and sample size is linearly proportional to the time each method takes (speed). As the results shown in Tables 3 and 4 with 200 sample size condition, the average speed of the grid search is around 120 seconds, which is almost twice the time spent for grid search with 100 samples shown in Tables 1 and 2. For the condition of the QTL location, it has no effect on the speed at all, as can be seen by comparing Tables 1, 3 and Tables 2, 4.

Tables 5-8 show the average speeds for F2 simulated data sets. We have almost the same conclusions for the F2 data sets as well as the BC data sets. The average speeds for F2 data sets are slower than the average speeds for the BC data sets, due to the fact that F2 has one more parameter to consider than BC. For example, the time required for the

grid search method of the F2 data sets with 100 samples is around 1000 seconds, which is almost 15 times slower compared to that of the BC data sets with 100 samples.

Table 1. Speed(unit:second) for BC when sample size = 100, QTL location = 10.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	60.3417	0.0232	0.0198	1.8375	3.6603
Marker Distance = 5 cM	59.6452	0.0274	0.0242	1.8207	3.6720
Marker Distance = 10 cM	59.3523	0.0308	0.0273	1.8110	3.8403
Marker Distance = 20 cM	59.2833	0.0362	0.0330	1.8194	3.6414
Heritability=0.05					
Marker Distance = 1 cM	60.3529	0.0238	0.0206	1.8315	3.6687
Marker Distance = 5 cM	59.6538	0.0274	0.0244	1.8218	3.6741
Marker Distance = 10 cM	59.3460	0.0313	0.0280	1.8120	3.8407
Marker Distance = 20 cM	59.3247	0.0371	0.0337	1.8245	3.6467
Heritability=0.1					
Marker Distance = 1 cM	60.1745	0.0241	0.0208	1.8344	3.6750
Marker Distance = 5 cM	59.7302	0.0280	0.0249	1.8283	3.6924
Marker Distance = 10 cM	59.3748	0.0316	0.0283	1.8136	3.8343
Marker Distance = 20 cM	59.4455	0.0377	0.0343	1.8267	3.6489
Heritability=0.15					
Marker Distance = 1 cM	60.2730	0.0243	0.0210	1.8323	3.6640
Marker Distance = 5 cM	59.7302	0.0280	0.0249	1.8283	3.6924
Marker Distance = 10 cM	59.3421	0.0322	0.0288	1.8131	3.8441
Marker Distance = 20 cM	61.0812	0.0393	0.0361	1.8220	3.6454

Table 2. Speed(unit:second) for BC when sample size = 100, QTL location = 45.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	60.3161	0.0237	0.0203	1.8347	3.6807
Marker Distance = 5 cM	59.6456	0.0274	0.0240	1.8184	3.6730
Marker Distance = 10 cM	59.3548	0.0435	0.0399	1.8242	3.8415
Marker Distance = 20 cM	59.4024	0.0365	0.0330	1.8267	3.6506
Heritability=0.05					
Marker Distance = 1 cM	60.3256	0.0239	0.0206	1.8335	3.6849
Marker Distance = 5 cM	59.6415	0.0277	0.0245	1.8212	3.6674
Marker Distance = 10 cM	59.5176	0.0313	0.0283	1.8189	3.8592
Marker Distance = 20 cM	59.3697	0.0374	0.0335	1.8202	3.6429
Heritability=0.1					
Marker Distance = 1 cM	60.3382	0.0242	0.0211	1.8346	3.6807
Marker Distance = 5 cM	59.6106	0.0280	0.0249	1.8228	3.6666
Marker Distance = 10 cM	59.3786	0.0317	0.0287	1.8149	3.8415
Marker Distance = 20 cM	59.6236	0.0381	0.0344	1.8213	3.6390
Heritability=0.15					
Marker Distance = 1 cM	60.3232	0.0244	0.0213	1.8307	3.6693
Marker Distance = 5 cM	59.7045	0.0284	0.0251	1.8211	3.6724
Marker Distance = 10 cM	59.3497	0.0323	0.0290	1.8153	3.8400
Marker Distance = 20 cM	59.5195	0.0387	0.0348	1.8300	3.6457

Table 3. Speed(unit:second) for BC when sample size = 200, QTL location = 10.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	118.6885	0.0474	0.0401	3.5897	7.8309
Marker Distance = 5 cM	117.8674	0.0536	0.0471	3.5720	7.1941
Marker Distance = 10 cM	117.6913	0.0597	0.0532	3.5754	7.2309
Marker Distance = 20 cM	117.3822	0.0715	0.0633	3.5798	9.4635
Heritability=0.05					
Marker Distance = 1 cM	118.5779	0.0481	0.0409	3.5946	7.8334
Marker Distance = 5 cM	117.9607	0.0547	0.0479	3.5815	7.2451
Marker Distance = 10 cM	117.6413	0.0611	0.0552	3.5714	7.2113
Marker Distance = 20 cM	117.4904	0.0730	0.0650	3.5820	7.1206
Heritability=0.1					
Marker Distance = 1 cM	118.5709	0.0505	0.0435	3.5910	7.8366
Marker Distance = 5 cM	117.8735	0.0551	0.0486	3.5756	7.2244
Marker Distance = 10 cM	117.6828	0.0620	0.0559	3.5721	7.2063
Marker Distance = 20 cM	117.9002	0.0740	0.0665	3.5832	7.1193
Heritability=0.15					
Marker Distance = 1 cM	118.6737	0.0490	0.0418	3.5993	7.8093
Marker Distance = 5 cM	117.9274	0.0560	0.0490	3.5750	7.3206
Marker Distance = 10 cM	117.6479	0.0632	0.0569	3.5785	7.2275
Marker Distance = 20 cM	117.5612	0.0751	0.0678	3.5910	7.1274

Table 4. Speed(unit:second) for BC when sample size = 200, QTL location = 45.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	118.5856	0.0469	0.0401	3.5870	7.8091
Marker Distance = 5 cM	118.1441	0.0536	0.0472	3.5785	7.2589
Marker Distance = 10 cM	117.8861	0.0602	0.0538	3.9692	7.2077
Marker Distance = 20 cM	117.4769	0.0707	0.0638	3.5848	7.1262
Heritability=0.05					
Marker Distance = 1 cM	118.5606	0.0480	0.0410	3.5844	7.8202
Marker Distance = 5 cM	123.3231	0.0560	0.0488	3.5774	7.2321
Marker Distance = 10 cM	117.5228	0.0618	0.0555	3.5676	7.2088
Marker Distance = 20 cM	117.7131	0.0728	0.0659	3.5864	7.4814
Heritability=0.1					
Marker Distance = 1 cM	118.8081	0.0488	0.0417	3.5936	7.8286
Marker Distance = 5 cM	117.8866	0.0563	0.0492	3.5757	7.2340
Marker Distance = 10 cM	117.6759	0.0631	0.0569	3.5788	7.2318
Marker Distance = 20 cM	117.4915	0.0753	0.0676	3.6003	7.1099
Heritability=0.15					
Marker Distance = 1 cM	118.7471	0.0492	0.0422	3.5920	7.8104
Marker Distance = 5 cM	117.8359	0.0569	0.0501	3.5786	7.2387
Marker Distance = 10 cM	117.5787	0.0643	0.0576	3.5740	7.2197
Marker Distance = 20 cM	117.3414	0.0768	0.0696	3.5782	7.1158

Table 5. Speed(unit:second) for F2 when sample size = 100, QTL location = 10.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	971.9750	0.2587	0.2531	6.7817	13.2640
Marker Distance = 5 cM	1015.0600	0.2363	0.2289	7.1350	13.3484
Marker Distance = 10 cM	977.1410	0.2471	0.2366	6.8407	13.4009
Marker Distance = 20 cM	932.5660	0.2341	0.2240	6.5401	13.9282
Heritability=0.05					
Marker Distance = 1 cM	1029.4230	0.2645	0.2587	6.8049	13.2723
Marker Distance = 5 cM	1006.6130	0.2412	0.2348	6.9888	13.3487
Marker Distance = 10 cM	977.6420	0.2529	0.2426	6.8480	13.4549
Marker Distance = 20 cM	930.1950	0.2379	0.2284	6.5140	13.9179
Heritability=0.1					
Marker Distance = 1 cM	972.0000	0.2711	0.2637	6.7859	13.2433
Marker Distance = 5 cM	1018.2260	0.2462	0.2404	7.0592	13.4161
Marker Distance = 10 cM	922.8560	0.2220	0.2129	6.4417	12.5376
Marker Distance = 20 cM	1093.1570	0.2426	0.2330	6.5158	13.9138
Heritability=0.15					
Marker Distance = 1 cM	972.0700	0.2786	0.2720	6.7808	13.2604
Marker Distance = 5 cM	1016.081	0.2518	0.2470	7.3087	13.3617
Marker Distance = 10 cM	922.2060	0.2249	0.2156	6.4384	12.5425
Marker Distance = 20 cM	929.6060	0.2488	0.2388	6.5159	13.9302

Table 6. Speed(unit:second) for F2 when sample size = 100, QTL location = 45.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	912.9090	0.2265	0.2209	6.3483	12.3652
Marker Distance = 5 cM	1016.2470	0.2359	0.2287	7.0842	13.3894
Marker Distance = 10 cM	976.5730	0.2475	0.2366	6.8418	13.4108
Marker Distance = 20 cM	930.4980	0.2338	0.2247	6.5226	13.9340
Heritability=0.05					
Marker Distance = 1 cM	971.1540	0.2653	0.2585	6.7783	13.2509
Marker Distance = 5 cM	1017.830	0.2409	0.2325	7.1235	13.3516
Marker Distance = 10 cM	977.0230	0.2538	0.2426	6.8401	13.3934
Marker Distance = 20 cM	930.0270	0.2376	0.2281	6.5242	13.9411
Heritability=0.1					
Marker Distance = 1 cM	914.5430	0.2388	0.2319	6.3728	12.3892
Marker Distance = 5 cM	1017.9810	0.2484	0.2391	7.0910	13.3989
Marker Distance = 10 cM	924.9740	0.2239	0.2129	6.4564	12.5682
Marker Distance = 20 cM	931.1610	0.2441	0.2358	6.5286	13.9523
Heritability=0.15					
Marker Distance = 1 cM	916.2490	0.2462	0.2384	6.3865	12.4278
Marker Distance = 5 cM	915.9400	0.2201	0.2111	6.3987	12.4448
Marker Distance = 10 cM	977.3440	0.2667	0.2550	6.8417	13.4236
Marker Distance = 20 cM	929.4420	0.2476	0.2409	6.5161	13.9590

Table 7. Speed(unit:second) for F2 when sample size = 200, QTL location = 10.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	1940.6900	0.4568	0.4381	13.5069	26.3629
Marker Distance = 5 cM	1830.2590	0.4716	0.4508	12.7354	24.8208
Marker Distance = 10 cM	2390.6850	0.4235	0.4074	14.4055	25.5853
Marker Distance = 20 cM	1871.2710	0.4500	0.4383	13.1206	25.2714
Heritability=0.05					
Marker Distance = 1 cM	1825.0680	0.3997	0.3849	12.6992	24.6970
Marker Distance = 5 cM	1999.1440	0.5531	0.5326	13.9325	26.5388
Marker Distance = 10 cM	2070.1010	0.4057	0.4153	14.4163	27.8382
Marker Distance = 20 cM	1866.4240	0.4567	0.4440	13.0913	25.2700
Heritability=0.1					
Marker Distance = 1 cM	1821.3510	0.4109	0.3962	12.6566	24.6470
Marker Distance = 5 cM	1830.3650	0.4925	0.4745	12.7854	24.8306
Marker Distance = 10 cM	2066.2860	0.4438	0.4273	14.3597	27.8501
Marker Distance = 20 cM	1864.7980	0.4672	0.4525	13.0736	25.2015
Heritability=0.15					
Marker Distance = 1 cM	1906.7430	0.4213	0.4077	12.6516	24.6735
Marker Distance = 5 cM	1830.2520	0.5060	0.4880	12.7377	24.7797
Marker Distance = 10 cM	2072.337	0.4566	0.4389	14.4215	27.9032
Marker Distance = 20 cM	1866.8880	0.4792	0.4643	13.1269	25.2514

Table 8. Speed(unit:second) for F2 when sample size = 200, QTL location = 45.1 cM

Methods	Grid Search	Laplace	Laplace fixed	IMS	MCMC
Heritability=0					
Marker Distance = 1 cM	1822.2490	0.3932	0.3762	12.6713	24.6592
Marker Distance = 5 cM	1832.0760	0.4725	0.4506	12.7386	24.7834
Marker Distance = 10 cM	2071.3200	0.4229	0.4074	14.4116	27.8797
Marker Distance = 20 cM	1873.647	0.4503	0.4369	13.1088	25.3048
Heritability=0.05					
Marker Distance = 1 cM	1904.4020	0.4029	0.3853	12.6806	24.7061
Marker Distance = 5 cM	1830.9660	0.4837	0.4602	12.7473	24.7868
Marker Distance = 10 cM	2069.1290	0.4347	0.4180	19.0286	27.8639
Marker Distance = 20 cM	1968.4140	0.5359	0.5236	13.8024	26.9577
Heritability=0.1					
Marker Distance = 1 cM	1827.2800	0.4131	0.3955	12.7035	24.7298
Marker Distance = 5 cM	1831.5750	0.4963	0.4747	12.7495	24.8071
Marker Distance = 10 cM	2070.4250	0.4444	0.4268	14.4784	27.9172
Marker Distance = 20 cM	1875.3520	0.4708	0.4552	13.1258	25.2686
Heritability=0.15					
Marker Distance = 1 cM	1820.4060	0.4235	0.4047	12.6720	24.6588
Marker Distance = 5 cM	1830.4810	0.5103	0.4887	12.8126	24.7868
Marker Distance = 10 cM	2165.3740	0.5259	0.5105	15.1095	29.1259
Marker Distance = 20 cM	1970.5640	0.5661	0.5485	13.8192	26.9454

Accuracy

As we have described, the numerical grid search integration is used here as the gold standard method. We compute average errors for all other methods using the formula (2.17) and the error potentially ranges from 0 to 1. If the error we obtain for one method is near 0, it means that the method we compared is almost as accurate as the gold standard grid search. But if the error for one method is near 1, this means that the posterior curve for the method we compared does not overlay at all with the gold standard posterior curve. We propose using the Laplace approximation method for detecting the location of the QTL so the error of the Laplace approximation method is used to evaluate how good this method compared to the gold standard method. In order to evaluate whether all other methods are more accurate than the Laplace approximation method, we report the ratios of the average errors for those methods to the average error for the Laplace approximation method. Any method with an error ratio less than 1 is considered to be more accurate than the Laplace approximation method.

Tables 9-16 display the average errors for the Laplace approximation method and the error ratios for the other methods. Tables 9-12 are the results for the simulated BC data sets and Tables 13-16 are the results for the simulated F2 data sets. For BC and F2, we generate data sets under 64 combinations of conditions, including choices of sample size, the QTL location, inter-marker distance (number of markers), and heritability effect, the same as those for the average speed evaluation.

Table 9 shows the results for a BC population with sample size equal to 100 and the QTL location is at 10.1 under different heritabilities and inter-marker distances. The Laplace approximation method has a very small average error (less than 1%) under all conditions, which indicates that the Laplace approximation method performs as well as the grid search method. All the other methods have error ratios greater than 1, which means that they are worse than the Laplace approximation method. Among them, the *Laplace fixed* approximation method is closest to the Laplace approximation method because the only difference between them is that the *Laplace fixed* approximation method uses the same information matrix for all possible QTL locations. The importance sampling and MCMC methods

perform much worse than the Laplace approximation method and the *Laplace fixed* approximation method. The inter-marker distance, i.e. the number of markers, doesn't have any significant effect on the accuracy. However, the heritability of the data sets has some effect. When sample size equals 200, we can see that larger heritability corresponds to higher accuracy, e.g. in Table 11, the average errors for the Laplace approximation method with the inter-marker distance 1 cM are 0.00057, 0.00199, 0.0017, 0.00047 for heritabilities 0, 0.05, 0.1, 0.15, respectively. These errors decrease as the heritability of the data sets increases. By comparing Table 9 and Table 10, we found that the location of the QTL doesn't have any significant influence on the accuracy. However, the number of the samples has a very significant effect on the accuracy (compare tables 9,10 with 11, 12). More samples correspond to higher accuracy.

Because the adaptive quadrature method takes a lot of time to run for the error calculation and does not show high accuracy, just one simulated result is listed for BC as an example: the error is 0.01192 when the inter-marker distance equals 10 cM, QTL position is at 45.1 cM, sample size is 200 and heritability is 0.15. This error is quite high compared to 0.00045, the error of the Laplace approximation method, so this method will no longer be considered here.

Table 13 to Table 16 report the average errors and error ratios for the F2 simulated data sets. The average errors in the F2 population are generally smaller than the average errors in the BC population. From the tables shown, the Laplace approximation method is always the best method among all the methods. (i.e. the error is always the smallest.) *The Laplace fixed* approximation method is the second best method. Its average error is just two to four times higher than the average error of the Laplace approximation method and its accuracy is much better than either the importance sampling or MCMC methods.

In summary, the Laplace approximation method and the *Laplace fixed* approximation method that we propose have very small errors and they are more than 1000 times faster than the grid search method. The *Laplace fixed* approximation method is a little faster than the Laplace approximation method, and has a moderate error compared to the Laplace approximation method so both of them can be considered in the eQTL analysis. All the other methods are slower and less accurate than the Laplace-related methods. Therefore,

the Laplace approximation method and the *Laplace fixed* approximation method are good replacements for the grid search method and the standard Bayesian QTL methods, given their accuracy and much faster speed.

Table 9. Error for BC when sample size = 100, QTL location = 10.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00107	2.67	3.03	12.52
Marker Distance = 5 cM	0.00092	2.49	3.11	7.63
Marker Distance = 10 cM	0.00102	2.70	3.57	6.60
Marker Distance = 20 cM	0.00166	2.56	2.75	3.54
Heritability=0.05				
Marker Distance = 1 cM	0.00345	2.78	3.34	18.88
Marker Distance = 5 cM	0.00272	3.05	3.38	11.17
Marker Distance = 10 cM	0.00309	3.00	4.01	8.98
Marker Distance = 20 cM	0.00296	2.45	3.40	5.65
Heritability=0.1				
Marker Distance = 1 cM	0.00504	3.02	3.54	21.84
Marker Distance = 5 cM	0.00391	3.36	3.60	15.29
Marker Distance = 10 cM	0.00392	3.35	4.24	13.29
Marker Distance = 20 cM	0.00451	2.37	3.41	7.62
Heritability=0.15				
Marker Distance = 1 cM	0.00399	4.33	4.14	36.49
Marker Distance = 5 cM	0.00519	3.26	3.34	15.65
Marker Distance = 10 cM	0.00379	3.96	3.75	12.36
Marker Distance = 20 cM	0.00537	2.56	3.58	6.92

Table 10. Error for BC when sample size = 100, QTL location = 45.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00106	2.61	2.97	11.90
Marker Distance = 5 cM	0.00083	2.47	3.22	7.13
Marker Distance = 10 cM	0.00109	2.82	3.55	6.43
Marker Distance = 20 cM	0.00123	2.93	3.39	4.21
Heritability=0.05				
Marker Distance = 1 cM	0.00373	2.90	3.05	15.37
Marker Distance = 5 cM	0.00313	2.74	3.44	9.99
Marker Distance = 10 cM	0.00341	2.85	3.82	8.42
Marker Distance = 20 cM	0.00314	2.60	3.06	4.72
Heritability=0.1				
Marker Distance = 1 cM	0.00451	3.53	3.72	24.56
Marker Distance = 5 cM	0.00450	3.29	3.76	12.28
Marker Distance = 10 cM	0.00452	3.11	3.55	9.77
Marker Distance = 20 cM	0.00477	2.74	3.50	5.07
Heritability=0.15				
Marker Distance = 1 cM	0.00413	4.34	3.56	31.87
Marker Distance = 5 cM	0.00352	4.69	4.37	17.89
Marker Distance = 10 cM	0.00305	4.74	3.96	17.57
Marker Distance = 20 cM	0.00494	3.20	3.58	6.92

Table 11. Error for BC when sample size = 200, QTL location = 10.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00057	2.06	6.50	19.77
Marker Distance = 5 cM	0.00074	2.07	6.13	20.85
Marker Distance = 10 cM	0.00057	2.93	9.32	14.03
Marker Distance = 20 cM	0.00064	3.47	7.39	13.13
Heritability=0.05				
Marker Distance = 1 cM	0.00199	3.18	9.88	66.69
Marker Distance = 5 cM	0.00277	2.58	7.61	29.99
Marker Distance = 10 cM	0.00323	2.74	6.70	29.89
Marker Distance = 20 cM	0.00269	2.96	6.77	15.19
Heritability=0.1				
Marker Distance = 1 cM	0.00170	5.09	9.62	127.50
Marker Distance = 5 cM	0.00199	4.29	9.27	78.82
Marker Distance = 10 cM	0.00200	4.53	7.88	54.06
Marker Distance = 20 cM	0.00301	3.28	7.94	24.96
Heritability=0.15				
Marker Distance = 1 cM	0.00047	17.38	22.75	510.04
Marker Distance = 5 cM	0.00096	8.51	10.99	156.46
Marker Distance = 10 cM	0.00036	21.26	24.01	312.58
Marker Distance = 20 cM	0.00091	8.33	14.37	61.55

Table 12. Error for BC when sample size = 200, QTL location = 45.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00055	2.04	6.65	22.15
Marker Distance = 5 cM	0.00074	2.07	6.11	14.48
Marker Distance = 10 cM	0.00053	2.79	9.83	13.78
Marker Distance = 20 cM	0.00058	3.87	7.63	11.35
Heritability=0.05				
Marker Distance = 1 cM	0.00223	3.24	8.76	42.74
Marker Distance = 5 cM	0.00282	2.64	7.58	19.78
Marker Distance = 10 cM	0.00390	2.25	5.59	13.99
Marker Distance = 20 cM	0.00286	3.23	8.05	13.77
Heritability=0.1				
Marker Distance = 1 cM	0.00214	4.61	9.28	57.95
Marker Distance = 5 cM	0.00208	4.71	8.97	39.82
Marker Distance = 10 cM	0.00188	4.56	9.63	35.83
Marker Distance = 20 cM	0.00179	6.20	10.45	28.19
Heritability=0.15				
Marker Distance = 1 cM	0.00027	29.16	35.21	435.80
Marker Distance = 5 cM	0.00100	9.04	13.48	66.38
Marker Distance = 10 cM	0.00045	15.46	28.08	140.65
Marker Distance = 20 cM	0.00102	9.25	13.34	49.51

Table 13. Error for F2 when sample size = 100, QTL location = 10.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00044	2.09	6.33	11.57
Marker Distance = 5 cM	0.00090	2.90	5.21	7.72
Marker Distance = 10 cM	0.00033	6.08	6.15	16.13
Marker Distance = 20 cM	0.00085	4.10	8.86	5.34
Heritability=0.05				
Marker Distance = 1 cM	0.00541	2.12	4.56	12.53
Marker Distance = 5 cM	0.00124	4.15	7.25	29.42
Marker Distance = 10 cM	0.00142	4.73	6.17	13.90
Marker Distance = 20 cM	0.00332	2.71	4.89	6.76
Heritability=0.1				
Marker Distance = 1 cM	0.00515	3.51	5.69	31.08
Marker Distance = 5 cM	0.00317	3.28	6.80	13.85
Marker Distance = 10 cM	0.00433	4.09	5.80	8.67
Marker Distance = 20 cM	0.00400	4.04	6.12	8.43
Heritability=0.15				
Marker Distance = 1 cM	0.00131	12.34	22.14	184.80
Marker Distance = 5 cM	0.00360	4.37	8.80	20.97
Marker Distance = 10 cM	0.00558	4.43	6.64	17.98
Marker Distance = 20 cM	0.01248	2.63	4.84	5.41

Table 14. Error for F2 when sample size = 100, QTL location = 45.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00046	2.32	6.21	10.14
Marker Distance = 5 cM	0.00218	2.23	3.39	7.48
Marker Distance = 10 cM	0.00040	3.70	5.50	9.60
Marker Distance = 20 cM	0.00128	2.44	2.67	2.07
Heritability=0.05				
Marker Distance = 1 cM	0.00581	1.87	3.92	7.72
Marker Distance = 5 cM	0.00147	4.96	5.19	51.02
Marker Distance = 10 cM	0.00074	5.63	5.58	10.66
Marker Distance = 20 cM	0.00167	4.56	4.86	7.98
Heritability=0.1				
Marker Distance = 1 cM	0.00439	3.75	11.09	28.94
Marker Distance = 5 cM	0.00250	4.22	5.86	47.85
Marker Distance = 10 cM	0.00273	3.85	6.68	5.55
Marker Distance = 20 cM	0.00272	5.77	10.90	9.78
Heritability=0.15				
Marker Distance = 1 cM	0.00191	11.96	25.29	96.54
Marker Distance = 5 cM	0.00291	5.17	9.62	36.35
Marker Distance = 10 cM	0.00808	2.37	4.83	4.11
Marker Distance = 20 cM	0.00386	5.72	10.15	12.70

Table 15. Error for F2 when sample size = 200, QTL location = 10.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00011	4.14	19.01	33.35
Marker Distance = 5 cM	0.00022	2.88	21.92	25.15
Marker Distance = 10 cM	0.00046	3.37	19.52	20.20
Marker Distance = 20 cM	0.00058	4.84	31.34	9.33
Heritability=0.05				
Marker Distance = 1 cM	0.00456	2.52	13.16	30.80
Marker Distance = 5 cM	0.01019	1.73	7.42	6.09
Marker Distance = 10 cM	0.00088	9.05	34.86	62.73
Marker Distance = 20 cM	0.00128	8.82	31.47	52.08
Heritability=0.1				
Marker Distance = 1 cM	0.00166	8.26	36.29	211.54
Marker Distance = 5 cM	0.00174	8.45	32.17	120.41
Marker Distance = 10 cM	0.00181	9.78	32.48	195.34
Marker Distance = 20 cM	0.00247	7.28	20.20	47.71
Heritability=0.15				
Marker Distance = 1 cM	0.00024	49.41	198.13	1137.33
Marker Distance = 5 cM	0.00073	24.38	54.07	374.78
Marker Distance = 10 cM	0.00187	11.70	27.83	183.72
Marker Distance = 20 cM	0.00382	5.15	16.25	45.41

Table 16. Error for F2 when sample size = 200, QTL location = 45.1 cM

$Methods$	$LaplaceError$	$\frac{Laplacefixed}{Laplace}$	$\frac{IMS}{Laplace}$	$\frac{MCMC}{Laplace}$
Heritability=0				
Marker Distance = 1 cM	0.00010	3.92	23.96	37.95
Marker Distance = 5 cM	0.00021	2.75	25.19	24.75
Marker Distance = 10 cM	0.00051	2.30	17.97	17.15
Marker Distance = 20 cM	0.00028	7.19	21.11	21.56
Heritability=0.05				
Marker Distance = 1 cM	0.00122	5.93	51.94	63.30
Marker Distance = 5 cM	0.00323	3.67	16.74	29.25
Marker Distance = 10 cM	0.00174	5.95	41.29	25.17
Marker Distance = 20 cM	0.00229	7.29	21.18	35.41
Heritability=0.1				
Marker Distance = 1 cM	0.00337	3.95	14.64	65.00
Marker Distance = 5 cM	0.00213	6.71	28.31	71.53
Marker Distance = 10 cM	0.00127	8.85	48.58	37.30
Marker Distance = 20 cM	0.00468	6.17	19.55	24.81
Heritability=0.15				
Marker Distance = 1 cM	0.00112	9.80	44.69	170.67
Marker Distance = 5 cM	0.00031	42.34	133.89	236.63
Marker Distance = 10 cM	0.00154	11.02	49.42	52.22
Marker Distance = 20 cM	0.00085	33.33	77.65	132.86

2.4 Real Data Analysis

We apply the proposed Bayesian approaches and all other existing methods to F2 real data from Ishimori *et al.* (2004). This F2 real data set was obtained from a cross of two highly divergent mouse strains: C57BL/6J (B6) mice (with low plasma HDL levels, and a susceptibility to atherosclerosis) and 129S1/SvImJ (129) mice (with high plasma HDL levels and some resistance to atherosclerosis). B6 males were mated to 129 females and their F1 progeny were intercrossed to produce 294 female F2 progeny. The results of Ishimori *et al.* (2004) suggest that there are some significant QTL linked to the phenotype plasma HDL cholesterol concentration. On chromosome 12, there is only one QTL linked to the HDL phenotype and this significant QTL has no interaction effect. We conducted the real data analysis on chromosome 12, because the assumption of the proposed one QTL model is that there is at most one QTL on the chromosome or chromosomes under study.

Chromosome 12 has 9 markers and is 66 cM long. 294 F2 mice were genotyped. If the genotype information for any markers is missing, the nearest non-missing genotype marker can be used as an alternative flanking marker. The phenotypic value HDL is log-transformed to follow the approximate Gaussian distribution. We generated 100 equally-spaced putative QTL locations across chromosome 12 to evaluate the posterior probabilities of the QTL at each location. We choose 0.5 as the prior for H_0 (no QTL is on this chromosome). The prior for the QTL at each location is specified as $\frac{0.5}{100}$.

Figure 2.5 shows the posterior probabilities of being a QTL for all putative QTL locations on chromosome 12. It also shows the posterior curve comparisons between the gold standard grid search method and all other methods. The top row and the first column plot illustrates the good fit of the Laplace approximation method and the grid search method. The posterior distribution from the *Laplace fixed* approximation method is also very close to that of the grid search method, shown on the top row, second column plot. For the second row, first column plot, the posterior points show the posterior distribution produced by the importance sampling method. Its posterior probability points show some variation compared with the posterior curve of the grid search method, but the peak locations of the posterior curves for the two methods are both detected at 20 cM. The second row,

second column plot shows the posterior probability points produced by the MCMC method compared to the posterior curve of the gold standard grid search method. The posterior probability points of the MCMC method shift its peak a little bit to the left compared to the posterior curve of the grid search method, but the peak locations of both methods detect still can be seen at around the same position, 20 cM.

In Figure 2.6, we take the difference for posterior probabilities between the method we evaluate and the gold standard method at each putative QTL location and draw this difference curve at all putative locations for the methods we evaluate. We found that the difference curve of the Laplace approximation is the smallest of all the methods. This result is consistent with the simulated result.

From the results of all the methods, we can say that the estimated QTL locations for all methods are all around 20 cM, which is the peak location of the posterior curves. Ishimori *et al.* (2004) also has reported exactly the same estimated QTL location - 20 cM. But for the posterior curves from the importance sampling and MCMC methods respectively, they have some variations compared to the posterior curve of the gold standard grid search method, and the speeds of the Laplace and *Laplace fixed* approximation methods are much faster than the speeds of the importance sampling and MCMC methods. Therefore, we can conclude again that the Laplace approximation and *Laplace fixed* approximation methods are the most accurate and have the highest speeds of all the methods in our real data analysis. The linkage posterior probability $p(H_A|data)$ is nearly 1 for the grid search method, and it is also nearly 1 for the Laplace approximation method. Because the linkage posterior probability is greater than 0.5, this supports the existence of the QTL on chromosome 12.

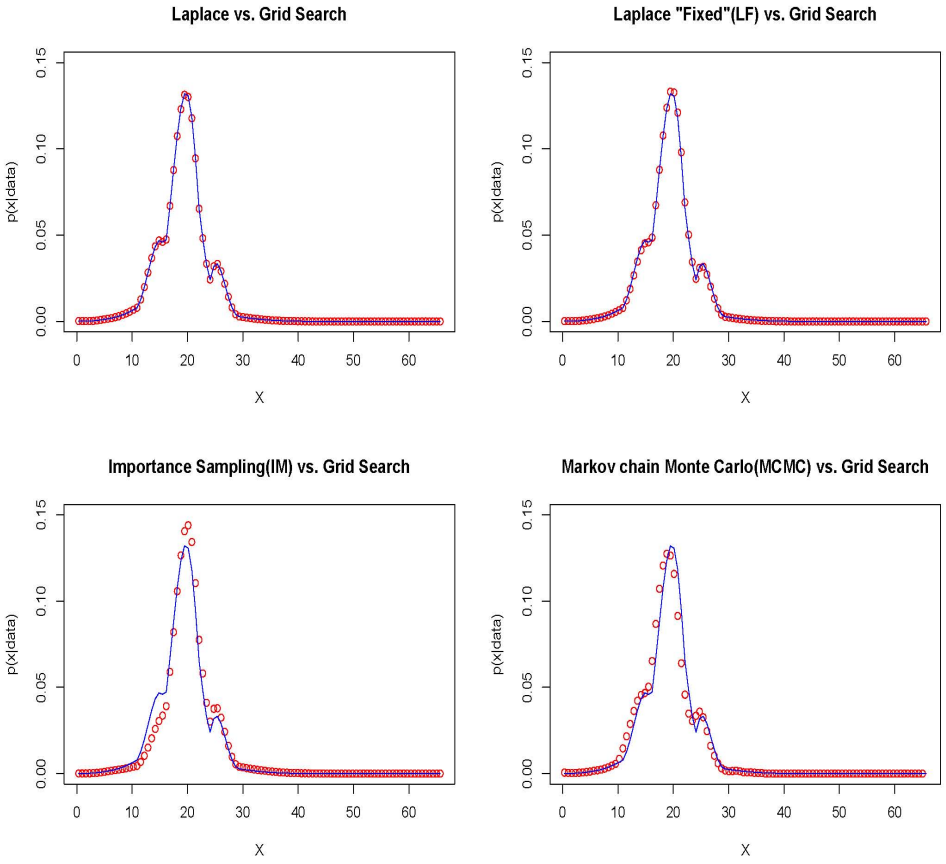

Figure 2.5: Posterior distributions of QTL locations for the chromosome 12 of the F2 data in Naoki et al. 2004. Several methods are applied and compared with the grid search method. The solid curves are for the grid search method, and the red scatter points are for the other methods.

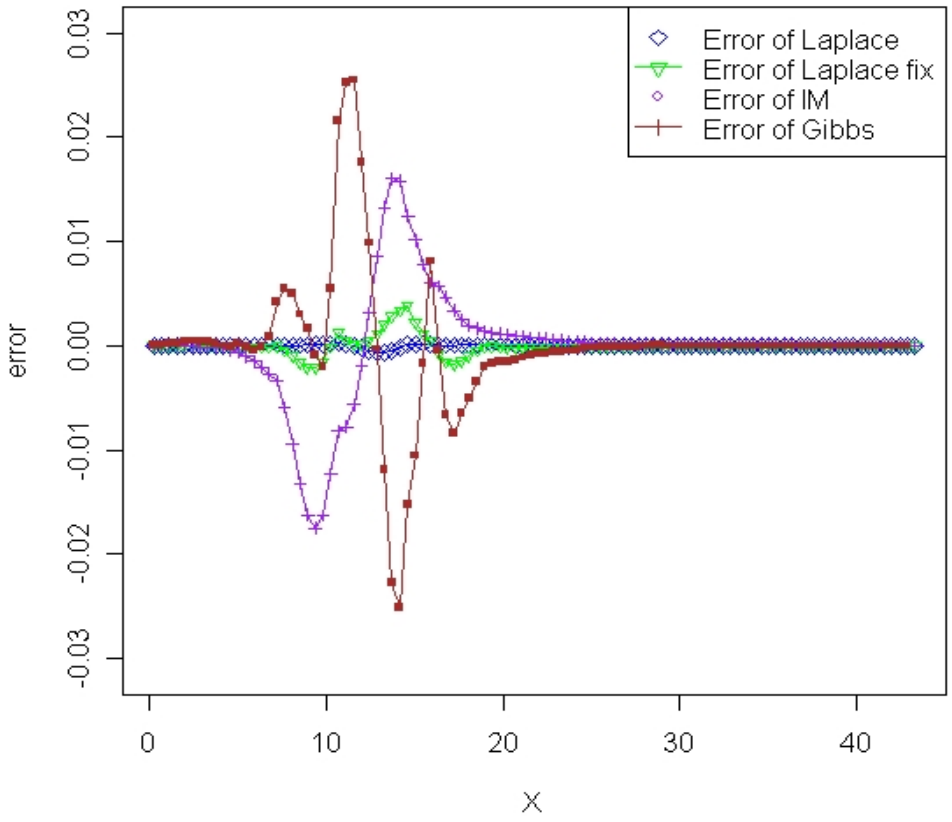

Figure 2.6: The difference of posterior probabilities from each method, compared with the grid search method along chromosome 12.

2.5 Application to eQTL analysis

We use the budding yeast data from Brem *et al.* (2002) to do the eQTL analysis. There are 112 yeast segregants from a cross of the two strains BY and RM, one haploid derivative. The operating characteristics of these data are essentially like that of a backcross, but with a higher effective recombination rate. The parent BY is a laboratory strain and the parent RM is a wild strain isolated from a California vineyard. Each segregant has 6229 gene expression traits with 2956 SNP markers. We treat each gene expression value on the Microarray data as a phenotype value. For each segregant, there are 6229 phenotypes we have to analyze, which is very computationally intensive. Because the data structure is high dimensional, our fast Bayesian QTL method via the Laplace approximation is applied in the eQTL analysis to save computation time.

We do the data management before analyzing the data. We delete the gene when over 20% of gene expression values are missing for all subjects, or when the information of gene location is missing, or when the information of which chromosome the gene belongs to is missing. Finally, 6139 genes are selected for eQTL analysis. If there are some missing gene expression values for the genes we select, we impute the missing gene expression values from the average of the existing gene expression values for all other subjects of that gene. For marker data, we delete the marker when there are over 20% missing genotype values for all subjects of that marker, or when the marker location information is missing, or when the information of which chromosome the marker belongs to is missing. We then still have 2956 markers in our eQTL analysis. When calculating the genotype probability of the likelihood function for all putative eQTL locations under the situation that the flanking markers are missing, we use the nearest existing marker genotype as the alternative flanking marker genotype.

In our data analysis, we use the equal prior probabilities of cis-acting, trans-acting, and unlinked for each transcript so the prior probability for each of them is 1/3. For each transcript (gene expression value), we use our Bayesian method to calculate the posterior probabilities for all putative eQTL locations and then subsequently summarized them into cis-acting posterior probability, trans-acting posterior probability, and unlinked posterior

probability. The cis-acting posterior probability is the posterior probability where the transcript whose gene expression is mapped to the gene location itself. The summation of the posterior probabilities for all other gene locations except the gene location itself is the trans-acting posterior probability. The unlinked posterior probability is calculated by using 1 - cis-acting posterior probability - trans-acting posterior probability. The maximum of these three posterior probabilities will be used to judge the linkage for that transcript. In Figure 2.7, we provide the eQTL analysis results for budding yeast. In this figure, the x axis represents the transcripts and the y axis is the gene locations. For one transcript, if the gene expression value is regulated by the gene location itself (cis-linked), we place a red dot at its gene location on the diagonal line of the plot; but if the gene expression value is regulated by other gene instead of the gene itself (trans-linked), there is a red dot at the gene location with the highest posterior probability which is at off diagonal line part. Among 6139 transcripts, we found that 23% are cis-linked genes and 31% are trans-linked genes. In our analysis, we also provide the information of the average posterior probability for all cis-acting genes, and for all trans-acting genes: the average posterior probability of all cis-acting genes is 0.2671, the average posterior probability of all trans-acting genes is 0.3773 and the average posterior probability of all unlinked genes is 0.3556.

We choose the gene with the highest cis-acting posterior, which is on chromosome 2, and the gene with the highest trans-acting posterior, which is also on chromosome 2, to draw the plot of its posterior probability against all gene locations (see Figure 2.8 and Figure 2.9, respectively). From these two figures, we can see the difference in the posterior probabilities between the cis-acting gene and trans-acting gene. In Figure 2.8, the posterior curve of the cis-acting gene has only one spike on the gene location itself, and the posterior probability is nearly 1 at that location. The posterior probabilities at other gene locations are very low compared to that at its gene location. In Figure 2.9, the posterior curve of the trans-acting gene has multiple peaks on several gene locations on chromosome 3, instead of having one peak posterior at its gene location on chromosome 2 compared to the posterior curve of the cis-acting gene. The posterior probability for gene location with the highest peak is 0.24 on chromosome 3. It is clear from these two figures that if the gene is cis-linked, then its gene expression value is regulated by the gene itself, but if the gene is trans-linked, then

the gene expression value is regulated by other gene instead of itself in the genome.

In Figure 2.7, we can see there are some apparent master control genes, which regulate many gene expression values. We find the top 12 master control genes using the somewhat criterion that if the gene regulates more than 30 gene expression values, then it is a master control gene. We list the top 12 master control genes in Table 17. We rank those genes from the most number of gene expressions controlled to the least number of gene expressions controlled. The first row shows the information for master control gene YOL082W on chromosome 15. It regulates 133 gene expression values and its gene location is from 168727 bp to 169974 bp. We also can see the information on other master control genes in this table. In Figure 2.7, several master control genes are very close to each other, e.g., gene YBR153W, gene YBR154C and, gene YBR156C are all on chromosome 2; gene YNL087W, gene YNL086W, gene YNL088W, gene YNL085W, and, gene YNL083W are all on chromosome 14. Therefore, we suspect those genes on the same chromosome are highly correlated and further statistical analysis is needed.

Our results for an eQTL analysis dataset serve as proof of principle that our Baesian approach is applicable to high throughput mapping problems. The high resolution and interpretability of our approach will enable straightforward refinement to (i) estimate cis vs. trans prior probability from the data. (ii) provide gene by gene analysis of linkage (because of our interval mapping), rather than marker by marker.

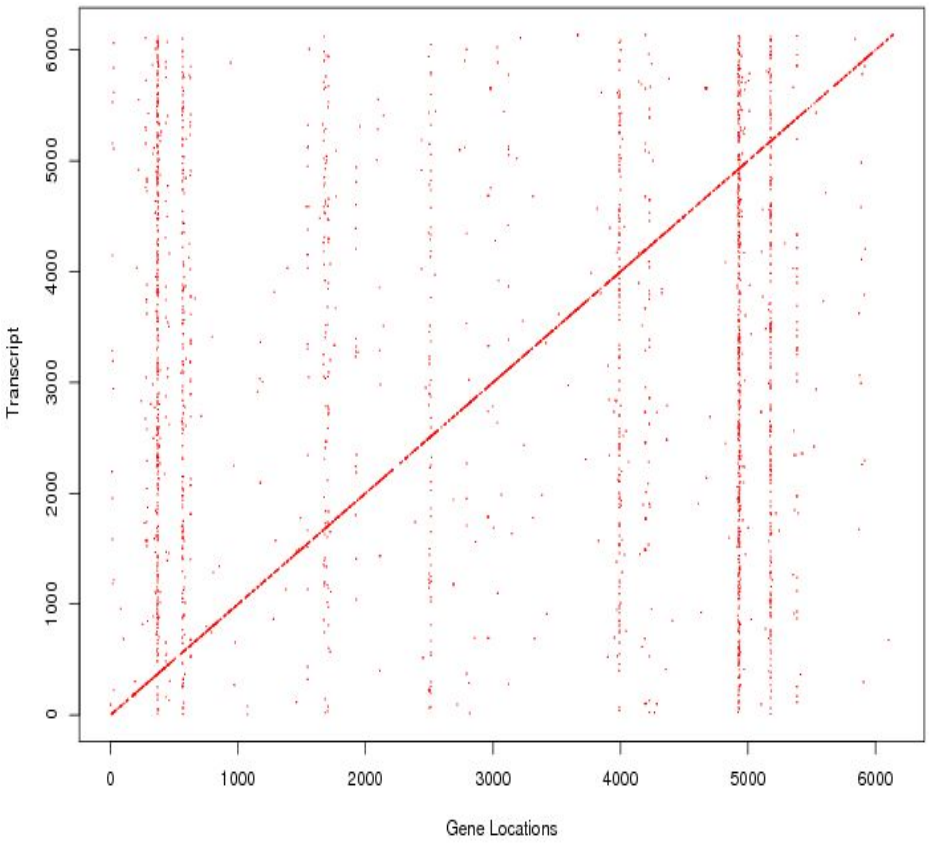

Figure 2.7: The eQTL plot for budding yeast data.

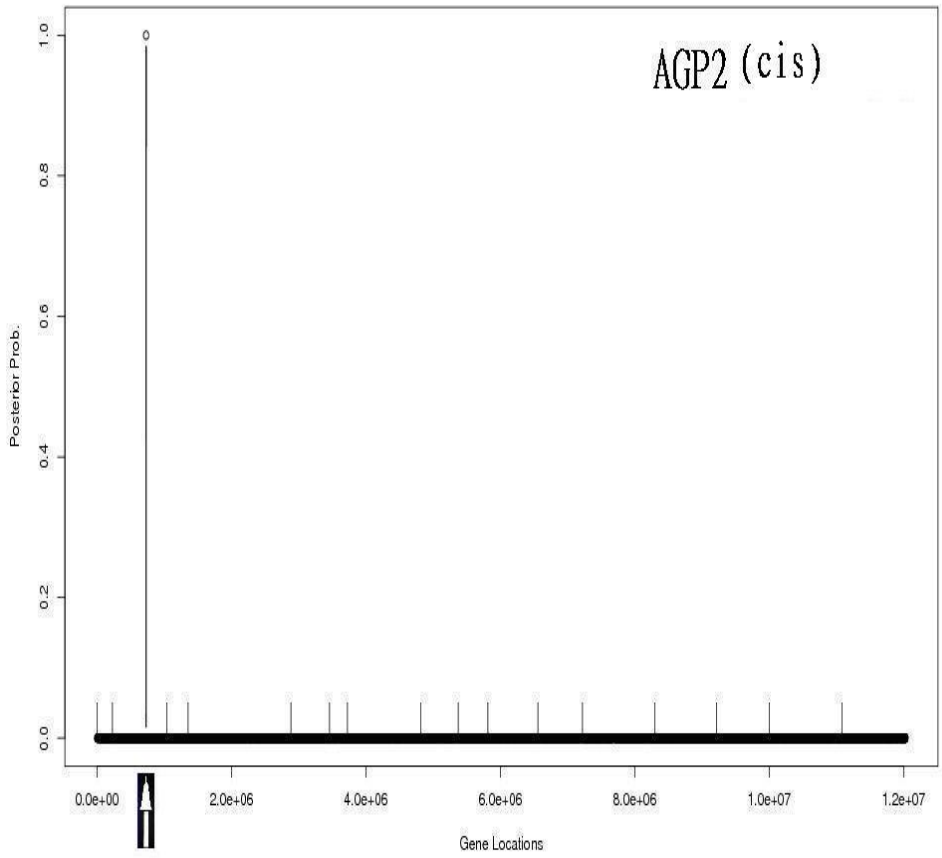

Figure 2.8: Posterior probability against all genome plot for transcript with the highest cis-acting posterior probability.

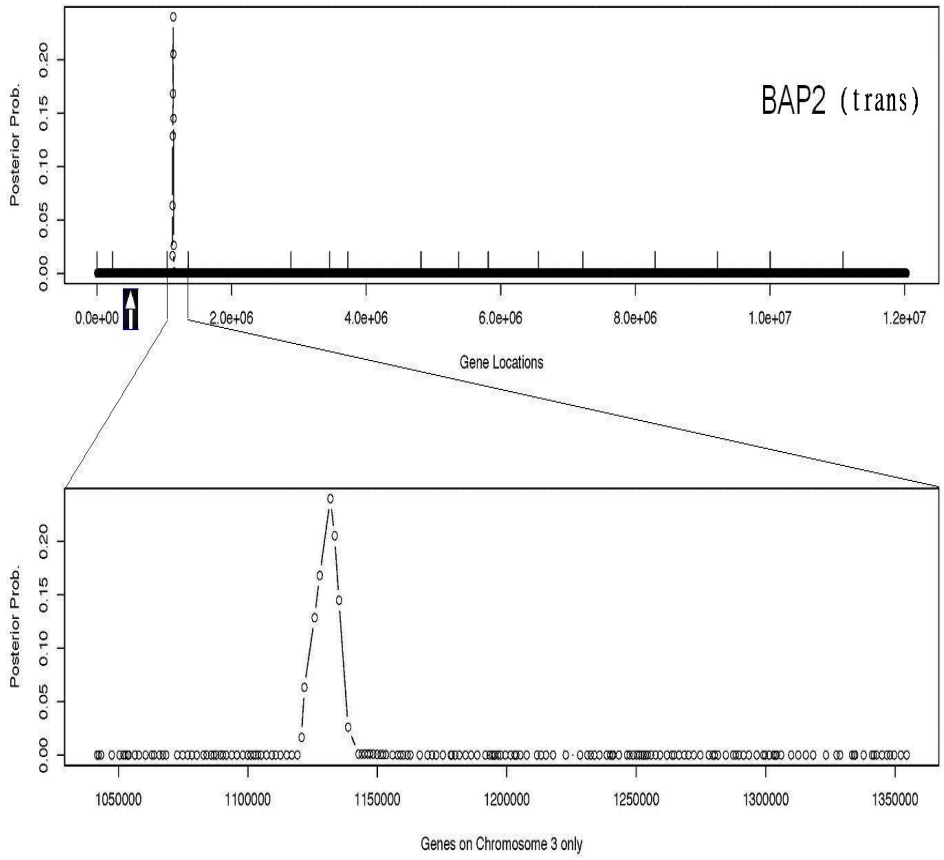

Figure 2.9: Posterior probability against all genome plot for transcript with the highest trans-acting posterior probability.

Table 17. Twelve putative master control genes in eQTL analysis

eQTL	*eQTL Location(b.p.)*	*chromosome*	*Number of gene eQTL controls*
YOL082W	168727, 169974	15	133
YBR153W	547454, 548188	2	99
YLR258W	660718, 662835	12	63
YNL087W	462413, 465949	14	54
YNL086W	466336, 466644	14	45
YNL088W	457706, 461992	14	40
YBR154C	549003, 548356	2	36
YCL025C	77919, 76018	3	35
YHR004C	113089, 111749	8	35
YBR156C	553194, 551098	2	32
YNL085W	467133, 469625	14	32
YNL083W	471379, 473016	14	30

2.6 Discussions

The Bayesian approaches we proposed—the Laplace approximation method and the *Laplace fixed* approximation method—are very fast compared to the competing methods. The average errors for both Laplace methods are also very small compared with the other methods examined here. As expected, the average error of the *Laplace fixed* approximation method is somewhat larger than the average error of the Laplace approximation method. However, while the increased error is modest, there is a several-fold speed improvement in QTL estimation. Therefore, our proposed approaches are good methods for detecting at most one QTL on the chromosomes under study and are suitable for high-throughput applications such as eQTL analyses, in which the location priors $p(x)$ are not necessarily uniform, because our method is allowed to specify the putative QTL location.

For the Bayesian approach we propose, we can directly get the linkage posterior probability $p(H_A|data)$ instead of using the Bayes factor to detect if there is a significant QTL. But we can still use the Bayes factor to judge the evidence for linkage. The following is the Bayes Factor formula for detecting linkage:

$$Bayes\ Factor = \frac{p(H_A|data)/p(H_A)}{p(H_0|data)/p(H_0)}.$$

The current Bayesian approaches we propose can only detect at most one QTL based on the chromosomes or all the genome under study, and the extension of the methods to enable detection of multiple QTL on the all genome is developed in the next Chapter. By using the proposed method, we also can provide insight into the connection between the LOD curve and the posterior probability for linkage. The applications of our method to eQTL analysis is a big step in this field because we have overcome the computation problem, and using this method, we can calculate the posterior probability directly, which is easy for us to interpret.

2.7 Appendix A: E-M algorithm for BC population in one QTL model

Suppose that we have n individuals in the BC population. For the i^{th} individual, the complete data set is (y_i, k), where y_i is the phenotype for the i^{th} individual. k is the 'missing' QTL genotype, equal to -1 if QTL genotype is Aa or 1 if QTL genotype is AA. We denote the flanking marker positions of QTL location x are $\{x_{left}, x_{right}\}$. The left flanking marker genotype of QTL is specified as $g(x_{left})$ and the right flanking marker genotype of QTL is specified as $g(x_{right})$ both equal to 1 or -1 depending on flanking marker genotypes. We are interested in the estimation of $\boldsymbol{\beta} = \{\mu_0, \mu_1, \sigma^2\}$, where $\mu_0 = \mu - a$ and $\mu_1 = \mu + a$ in Chapter 2. Therefore, E-M algorithm is applied to obtain $\boldsymbol{\beta}$. Let $\boldsymbol{\alpha} = (g(x_{left}), g(x_{right}), \boldsymbol{\beta})$.

$$\begin{aligned} f(y_i, k|\boldsymbol{\alpha}, x) &= \sum_{j=0}^{1} f(y_i|k = 2j-1, \boldsymbol{\alpha}, x) p(k = 2j-1|\boldsymbol{\alpha}, x) I(k = 2j-1) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \{exp(-\frac{(y_i - \mu + a)^2}{2\sigma^2}) g_0(x) I(k = -1) + exp(-\frac{(y_i - \mu - a)^2}{2\sigma^2}) g_1(x) I(k = 1)\} \end{aligned}$$

, where $g_0(x) = p(k = -1|\boldsymbol{\alpha}, x)$, $g_1(x) = p(k = 1|\boldsymbol{\alpha}, x)$, and $I(.)$ is the index function for QTL genotype.

Therefore,

$$f(\boldsymbol{y}|\boldsymbol{\alpha}, x) = \prod_{i=1}^{n} f(y_i, k|\boldsymbol{\alpha}, x)$$

We take log on both side of the equation and get

$$\begin{aligned} l &= ln(f(\boldsymbol{y}|\boldsymbol{\alpha}, x)) \\ &= \sum_{i=1}^{n} \{ln[\sum_{j=0}^{1} f(y_i|k = 2j-1, \boldsymbol{\alpha}, x) p(k = 2j-1|\boldsymbol{\alpha}, x) I(k = 2j-1)]\} \end{aligned}$$

In the E-step, we take the expectation for l given $\boldsymbol{y}, \boldsymbol{\alpha}_m$:

$$E[l|\boldsymbol{y},\boldsymbol{\alpha}_m]=\sum_{i=1}^{n}\{\sum_{j=0}^{1}ln[f(y_i|k=2j-1,\boldsymbol{\alpha},x)p(k=2j-1|\boldsymbol{\alpha},x)p(k=2j-1|y_i,\alpha_{\mathbf{m}})]\}$$

, where we denote $W_m(i,j)=p(k=j|y_i,\alpha_{\mathbf{m}})=\frac{\mathbf{f(y_i|k=2j-1,\boldsymbol{\alpha}_m,x)p(k=2j-1|\boldsymbol{\alpha}_m,x)}}{\sum_{\mathbf{j=0}}^{\mathbf{1}}\mathbf{f(y_i|k=2j-1,\boldsymbol{\alpha}_m,x)p(k=2j-1|\boldsymbol{\alpha}_m,x)}}$

In M step, we take the derivative of $E[l|\mathbf{y},\alpha_{\mathbf{m}}]$ with respective to $\mu_0,\mu 1,\sigma^2$,

$$\frac{\partial E[l|\mathbf{y},\alpha_{\mathbf{m}}]}{\partial\mu_p}=\frac{\sum_{i=1}^{n}(y_i-\mu_p)W_m(i,p)}{\sigma^2}=0$$

$$\Longrightarrow\hat{\mu}_p=\frac{\sum_{i=1}^{n}y_i\ W_m(i,p)}{\sum_{i=1}^{n}W_m(i,p)}$$

, where $p=0,1$.

$$\frac{\partial E[l|\mathbf{y},\alpha_{\mathbf{m}}]}{\partial\sigma^2}=\sum_{i=1}^{n}\sum_{p=0}^{1}(\frac{-1}{2\sigma^2}+\frac{(y_i-\mu_p)^2}{2(\sigma^2)^2})W_m(i,j)=0$$

$$\begin{aligned}\Longrightarrow\hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}\sum_{p=0}^{1}(y_i-\mu_p)^2\ W_m(i,j)}{\sum_{i=1}^{n}\sum_{p=0}^{1}W_m(i,j)}\\ &= \frac{\sum_{i=1}^{n}\sum_{p=0}^{1}(y_i-\mu_p)^2\ W_m(i,j)}{n}\end{aligned}$$

After we obtain the estimates of μ_0,μ_1,σ^2, we can easily obtain the estimates of μ,a,σ^2 by linear transformation. Also, the E-M algorithm of F2 population in one QTL model, E-M algorithm of BC population in two QTL model as well as E-M algorithm of BC population in eQTL analysis can be easily extended by using similar steps above.

2.8 Appendix B: Fisher Information Matrix Derivation under H_A for Backcross in One QTL Model

The alternative hypothesis in this section is that the location of QTL is on the chromosome under study and nuisance parameters are $\boldsymbol{\beta} = \{\mu,\ a\}$. Likelihood function for the i_{th} individual under the alternative hypothesis can be expressed as:

$$\begin{aligned} f(y_i|\boldsymbol{\beta}, x) &= \sum_{j=0}^{1} f(y_i|k = 2j-1, \boldsymbol{\beta}, x)p(k = 2j-1|\boldsymbol{\beta}, x) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\{exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})g_0(x) + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})g_1(x)\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\mathbf{A}_i, \end{aligned}$$

where

$$g_0(x) = p(k = -1|\boldsymbol{\beta}, x),\ g_1(x) = p(k = 1|\boldsymbol{\beta}, x),$$

$$\mathbf{A}_i = exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})g_0(x) + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})g_1(x),\ i = 1, \cdots, n.$$

The likelihood for all the individuals are :

$$f(\boldsymbol{y}|\boldsymbol{\beta}, x) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, x) = (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \mathbf{A}_i$$

After taking log for the likelihood above:

$$\ell = \ln f(\boldsymbol{y}|\boldsymbol{\beta}, x) \propto \frac{-n}{2}\ln\sigma^2 + \sum_{i=1}^{n} \ln \mathbf{A}_i$$

2.8.1 The First Derivatives of the Loglikelihood Function

We define

$$\mathbf{B0}_i \doteq \frac{\partial \mathbf{A}_i}{\partial \mu} = g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)}{\sigma^2} + g_1(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)}{\sigma^2}$$

$$\mathbf{B1}_i \doteq \frac{\partial \mathbf{A}_i}{\partial a} = -g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)}{\sigma^2}+g_1(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)}{\sigma^2}$$

and

$$\begin{aligned}\mathbf{B2}_i &\doteq \frac{\partial \mathbf{A}_i}{\partial \sigma^2} \\ &= g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)^2}{2(\sigma^2)^2} + g_1(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)^2}{2(\sigma^2)^2}\end{aligned}$$

Then,

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{\mathbf{B0}_i}{\mathbf{A}_i}$$

$$\frac{\partial \ell}{\partial a} = \sum_{i=1}^{n} \frac{\mathbf{B1}_i}{\mathbf{A}_i}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{\mathbf{B2}_i}{\mathbf{A}_i}$$

2.8.2 The Second Derivatives of the Loglikelihood Function

We define the following notations:

$$\begin{aligned}\mathbf{B00}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial \mu^2} = \frac{\partial \mathbf{B0}_i}{\partial \mu} \\ &= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[\frac{(y_i-\mu+a)^2}{\sigma^2}-1] \\ &+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]\end{aligned}$$

$$\begin{aligned}\mathbf{B01}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial \mu \partial a} = \frac{\partial \mathbf{B0}_i}{\partial a} = -exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[\frac{(y_i-\mu+a)^2}{\sigma^2}-1] \\ &+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]\end{aligned}$$

$$
\begin{aligned}
\mathbf{B02}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial\mu\partial\sigma^2} = \frac{\partial \mathbf{B2}_i}{\partial\mu} \\
&= g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)}{(\sigma^2)^2}[\frac{(y_i-\mu+a)^2}{2\sigma^2}-1] \\
&+ g_1(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{2\sigma^2}-1]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{B11}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial a^2} = \frac{\partial \mathbf{B1}_i}{\partial a} = exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[\frac{(y_i-\mu+a)^2}{\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{B12}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial a\partial\sigma^2} = \frac{\partial \mathbf{B2}_i}{\partial a} = -exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{(\sigma^2)^2}(y_i-\mu+a)[\frac{(y_i-\mu+a)^2}{2\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_1(x)}{(\sigma^2)^2}(y_i-\mu-a)[\frac{(y_i-\mu-a)^2}{2\sigma^2}-1]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{B22}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial(\sigma^2)^2} = \frac{\partial \mathbf{B2}_i}{\partial\sigma^2} \\
&= g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)^2}{(\sigma^2)^3}[\frac{(y_i-\mu+a)^2}{4\sigma^2}-1] \\
&+ g_1(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)^2}{\sigma^3}[\frac{(y_i-\mu-a)^2}{4\sigma^2}-1]
\end{aligned}
$$

Then, the second derivatives are

$$\frac{\partial^2 \ell}{\partial\mu^2} = \sum_{i=1}^{n}(\frac{\mathbf{B00}_i}{\mathbf{A}_i} - \frac{\mathbf{B0}_i^2}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial a^2} = \sum_{i=1}^{n}(\frac{\mathbf{B11}_i}{\mathbf{A}_i} - \frac{\mathbf{B1}_i^2}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial \mu \partial a} = \sum_{i=1}^{n} (\frac{\mathbf{B01}_i}{\mathbf{A}_i} - \frac{\mathbf{B0}_i \mathbf{B1}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \sum_{i=1}^{n} (\frac{\mathbf{B02}_i}{\mathbf{A}_i} - \frac{\mathbf{B2}_i \mathbf{B0}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial a \partial \sigma^2} = \sum_{i=1}^{n} (\frac{\mathbf{B12}_i}{\mathbf{A}_i} - \frac{\mathbf{B2}_i \mathbf{B1}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} + \sum_{i=1}^{n} (\frac{\mathbf{B22}_i}{\mathbf{A}_i} - \frac{\mathbf{B2}_i^2}{\mathbf{A}_i^2})$$

2.9 Appendix C: Fisher Information Matrix Derivation under H_A for F2 in One QTL Model

The alternative hypothesis in this section is that the location of QTL is on the chromosome under study and and nuisance parameters are $\boldsymbol{\beta} = \{\mu,\ a,\ d\}$. Likelihood function for the i_{th} individual under the alternative hypothesis can be expressed as:

$$\begin{aligned} f(y_i|\boldsymbol{\beta}, x) &= \sum_{j=0}^{2} f(y_i|k=j, \boldsymbol{\beta}, x) p(k=j|\boldsymbol{\beta}, x) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\{exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})g_0(x) + exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})g_1(x) \\ &\quad + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})g_2(x)\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\mathbf{A}_i, \end{aligned}$$

where

$$g_0(x) = p(k=0|\boldsymbol{\beta}, x),\ g_1(x) = p(k=1|\boldsymbol{\beta}, x),\ g_2(x) = p(k=2|\boldsymbol{\beta}, x),$$

$$\mathbf{A}_i = exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})g_0(x) + exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})g_1(x) + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})g_2(x),$$

$$and\ i = 1, \cdots, n.$$

The likelihood for all the individuals are :

$$f(\boldsymbol{y}|\boldsymbol{\beta}, x) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, x) = (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \mathbf{A}_i$$

After taking log for the likelihood above:

$$\ell = \ln f(\boldsymbol{y}|\boldsymbol{\beta}, x) \propto \frac{-n}{2}\ln\sigma^2 + \sum_{i=1}^{n} \ln \mathbf{A}_i$$

2.9.1 The First Derivatives of the Loglikelihood Function

We define

$$\begin{aligned}\mathbf{F0}_i &\doteq \frac{\partial \mathbf{A}_i}{\partial \mu} = g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)}{\sigma^2} + g_1(x)exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{(y_i-\mu-d)}{\sigma^2} \\ &+ g_2(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)}{\sigma^2}\end{aligned}$$

$$\mathbf{F1}_i \doteq \frac{\partial \mathbf{A}_i}{\partial a} = -g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)}{\sigma^2} + g_2(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)}{\sigma^2}$$

$$\mathbf{F2}_i \doteq \frac{\partial \mathbf{A}_i}{\partial d} = g_1(x)exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{(y_i-\mu-d)}{\sigma^2}$$

and

$$\begin{aligned}\mathbf{F3}_i &\doteq \frac{\partial \mathbf{A}_i}{\partial \sigma^2} = g_0(x)exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{(y_i-\mu+a)^2}{2(\sigma^2)^2} + g_1(x)exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{(y_i-\mu-d)^2}{2(\sigma^2)^2} \\ &+ g_2(x)exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{(y_i-\mu-a)^2}{2(\sigma^2)^2}\end{aligned}$$

Then,

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{\mathbf{F0}_i}{\mathbf{A}_i}$$

$$\frac{\partial \ell}{\partial a} = \sum_{i=1}^{n} \frac{\mathbf{F1}_i}{\mathbf{A}_i}$$

$$\frac{\partial \ell}{\partial d} = \sum_{i=1}^{n} \frac{\mathbf{F2}_i}{\mathbf{A}_i}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{\mathbf{F3}_i}{\mathbf{A}_i}$$

2.9.2 The Second Derivatives of the Loglikelihood Function

We define the following notations:

$$\mathbf{F00}_i \doteq \frac{\partial^2 \mathbf{A}_i}{\partial \mu^2} = \frac{\partial \mathbf{F0}_i}{\partial \mu}$$

$$
\begin{aligned}
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[\frac{(y_i-\mu+a)^2}{\sigma^2}-1] + exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-d)^2}{\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{F01}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial\mu\partial a} = \frac{\partial \mathbf{F0}_i}{\partial a} \\
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[-\frac{(y_i-\mu+a)^2}{\sigma^2}+1] + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]
\end{aligned}
$$

$$
\mathbf{F02}_i \doteq \frac{\partial^2 \mathbf{A}_i}{\partial\mu\partial d} = \frac{\partial \mathbf{F0}_i}{\partial d} = exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-d)^2}{\sigma^2}-1]
$$

$$
\begin{aligned}
\mathbf{F03}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial\mu\partial\sigma^2} = \frac{\partial \mathbf{F0}_i}{\partial\sigma^2} \\
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)(y_i-\mu+a)}{(\sigma^2)^2}[\frac{(y_i-\mu+a)^2}{2\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)(y_i-\mu-d)}{(\sigma^2)^2}[\frac{(y_i-\mu-d)^2}{2\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)(y_i-\mu-a)}{(\sigma^2)^2}[\frac{(y_i-\mu-a)^2}{2\sigma^2}-1]
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{F11}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial a^2} = \frac{\partial \mathbf{F1}_i}{\partial a} \\
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)}{\sigma^2}[\frac{(y_i-\mu+a)^2}{\sigma^2}-1] + exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)}{\sigma^2}[\frac{(y_i-\mu-a)^2}{\sigma^2}-1]
\end{aligned}
$$

$$
\mathbf{F12}_i \doteq \frac{\partial^2 \mathbf{A}_i}{\partial a\partial d} = \frac{\partial \mathbf{F1}_i}{\partial d} = 0
$$

$$
\begin{aligned}
\mathbf{F13}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial a \partial \sigma^2} = \frac{\partial \mathbf{F3}_i}{\partial a} \\
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)(y_i-\mu+a)}{(\sigma^2)^2}[-\frac{(y_i-\mu+a)^2}{2\sigma^2}+1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)(y_i-\mu-a)}{(\sigma^2)^2}[\frac{(y_i-\mu-a)^2}{2\sigma^2}-1]
\end{aligned}
$$

$$
\mathbf{F22}_i \doteq \frac{\partial^2 \mathbf{A}_i}{\partial d^2} = \frac{\partial \mathbf{F2}_i}{\partial d} = exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-d)^2}{\sigma^2}-1]
$$

$$
\mathbf{F23}_i \doteq \frac{\partial^2 \mathbf{A}_i}{\partial d \partial \sigma^2} = \frac{\partial \mathbf{F2}_i}{\partial \sigma^2} = exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)}{\sigma^2}[\frac{(y_i-\mu-d)^2}{2(\sigma^2)^2}-1](y_i-\mu-d)
$$

$$
\begin{aligned}
\mathbf{F33}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial (\sigma^2)^2} = \frac{\partial \mathbf{F3}_i}{\partial \sigma^2} \\
&= exp(-\frac{(y_i-\mu+a)^2}{2\sigma^2})\frac{g_0(x)(y_i-\mu+a)^2}{(\sigma^2)^3}[\frac{(y_i-\mu+a)^2}{4\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-d)^2}{2\sigma^2})\frac{g_1(x)(y_i-\mu-d)^2}{(\sigma^2)^3}[\frac{(y_i-\mu-d)^2}{4\sigma^2}-1] \\
&+ exp(-\frac{(y_i-\mu-a)^2}{2\sigma^2})\frac{g_2(x)(y_i-\mu-a)^2}{(\sigma^2)^3}[\frac{(y_i-\mu-a)^2}{4\sigma^2}-1]
\end{aligned}
$$

Then, the second derivatives are

$$
\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{i=1}^{n}(\frac{\mathbf{F00}_i}{\mathbf{A}_i} - \frac{\mathbf{F0}_i^2}{\mathbf{A}_i^2})
$$

$$
\frac{\partial^2 \ell}{\partial a^2} = \sum_{i=1}^{n}(\frac{\mathbf{F11}_i}{\mathbf{A}_i} - \frac{\mathbf{F1}_i^2}{\mathbf{A}_i^2})
$$

$$
\frac{\partial^2 \ell}{\partial d^2} = \sum_{i=1}^{n}(\frac{\mathbf{F22}_i}{\mathbf{A}_i} - \frac{\mathbf{F2}_i^2}{\mathbf{A}_i^2})
$$

$$\frac{\partial^2 \ell}{\partial \mu \partial a} = \sum_{i=1}^{n} (\frac{\mathbf{F01}_i}{\mathbf{A}_i} - \frac{\mathbf{F0}_i \mathbf{F1}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial \mu \partial d} = \sum_{i=1}^{n} (\frac{\mathbf{F02}_i}{\mathbf{A}_i} - \frac{\mathbf{F0}_i \mathbf{F2}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial a \partial d} = \sum_{i=1}^{n} (-\frac{\mathbf{F1}_i \mathbf{F2}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \sum_{i=1}^{n} (\frac{\mathbf{F03}_i}{\mathbf{A}_i} - \frac{\mathbf{F0}_i \mathbf{F3}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial a \partial \sigma^2} = \sum_{i=1}^{n} (\frac{\mathbf{F13}_i}{\mathbf{A}_i} - \frac{\mathbf{F3}_i \mathbf{F1}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial d \partial \sigma^2} = \sum_{i=1}^{n} (\frac{\mathbf{F23}_i}{\mathbf{A}_i} - \frac{\mathbf{F2}_i \mathbf{F3}_i}{\mathbf{A}_i^2})$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} + \sum_{i=1}^{n} (\frac{\mathbf{F33}_i}{\mathbf{A}_i} - \frac{\mathbf{F3}_i^2}{\mathbf{A}_i^2})$$

2.10 Appendix D: Derivation of MCMC Method for Backcross in One QTL Model

Suppose that we have n individuals in the Backcross population. Denote that $Y = (y_1, \cdots, y_n)$, where y_i is the phenotype of the ith individual. The genotype of the QTL for the ith individual is denoted as g_i, where $g_i = -1$ if the QTL genotype is Aa or $g_i = 1$ if the QTL genotype is AA. We use $G = (g_1, \cdots, g_n)$ to represent the QTL genotypes of all individuals. The objective is to find the location of QTL as well as the effect of QTL on the phenotype, which is measured by μ_0, μ_1, where $\mu_0 = \mu - a$ and $\mu_1 = \mu + a$ in Chapter 2.

For the i^{th} individual, the distribution of y_i is assumed to follow normal distribution with variance σ^2. The mean value is μ_0 if the QTL genotype is Aa, or μ_1 if the QTL genotype is AA. To find the QTL, markers are measured across the the chromosome. The locations and the genotypes of the markers are presented by M. Among all these parameters, $\{\mu_0, \mu_1, \sigma^2, G, x\}$ are unknown, and the phenotypes Y and the marker information M are known.

Here assuming the unknown parameters have the following prior distributions:

$$\begin{aligned} \mu_0 &\sim Unif(-s, s), \\ \mu_1 &\sim Unif(-s, s), \\ \sigma^2 &\sim Unif(0, c), \end{aligned} \tag{2.18}$$

where s and c are given constants.

Now we derive the conditional distributions for the nuisance parameters $\mu_0, \mu_1, \sigma^2, x, g_i$ sequatially.

Step 1: Sampling μ_0, μ_1

$$\begin{aligned} p(\mu_0|\mu_1, \sigma^2, x, G, Y, M) &\propto p(Y|\mu_0, \mu_1, \sigma^2, x, G, M)p(\mu_0, \mu_1, \sigma^2, x, G|M) \\ &\propto p(Y|\mu_0, \mu_1, \sigma^2, G)p(\mu_0) \\ &\propto exp\{-\frac{\sum_{i:g_i(x)=-1}(y_i - \mu_0)^2}{2\sigma^2}\} * I_{\{-s<\mu_0<s\}} \end{aligned} \tag{2.19}$$

Define n_0 and n_1 are the numbers of individuals whose QTL genotype is -1 or 1 respectively. And define $\bar{y_0} = \frac{1}{n_0}\sum_{i:g_i=-1} y_i$ and $\bar{y_0} = \frac{1}{n_1}\sum_{i:g_i=1} y_i$. The conditional distribution in Equation (2.19) is

$$\begin{aligned} p(\mu_0|\mu_1,\sigma^2,x,G,Y,M) &\propto exp\{-\frac{n_0(\mu_0-\bar{y_0})^2}{2\sigma^2}\} * I_{-s<\mu_0<s} \\ &\propto N(\bar{y_0},\sigma^2/n_0) * I_{-s<\mu_0<s} \end{aligned} \tag{2.20}$$

Thus the conditional distribution of μ_0 given other parameters, follows truncated Gaussian distribution. Similar result holds for μ_1:

$$p(\mu_1|\mu_0,\sigma^2,x,G,Y,M) \propto N(\bar{y_1},\sigma^2/n_1) * I_{-s<\mu_1<s}$$

Step 2: Sampling σ^2

$$\begin{aligned} p(\sigma^2|\mu_0,\mu_1,x,G,Y,M) &\propto p(Y|\mu_0,\mu_1,\sigma^2,x,G,M)p(\mu_0,\mu_1,\sigma^2,x,G|M) \\ &\propto p(Y|\mu_0,\mu_1,\sigma^2,G)p(\sigma^2) \\ &\propto (\frac{1}{\sqrt{\sigma^2}})^n exp\{-\frac{\sum_{i=1}^n (y_i-\mu_{g_i})^2}{2\sigma^2}\} * I_{0<\sigma^2<c}, \end{aligned}$$

which is truncated inverse $Gamma(\frac{n}{2}, \frac{\sum_{i=1}^n (y_i-\mu_{g_i})^2}{2})$. Hence $\frac{\sigma^2}{\sum_{i=1}^n (y_i-\mu_{g_i})^2} \sim$ truncated inverse $chisq(n)$.

Step 3: To Sample G, we sample each g_i separately.

$$\begin{aligned} p(g_i|\mu_0,\mu_1,\sigma^2,x,Y,M) &\propto f(y_i|\mu_0,\mu_1,\sigma^2,x,M)p(\mu_0,\mu_1,\sigma^2,x,g_i|M) \\ &\propto exp\{-\frac{(y_i-\mu_{g_i})^2}{2\sigma^2}\}p(g_i|x,M) \end{aligned}$$

Thus the condition distribution of $g_i = 0$ is

$$p(g_i = 0|\mu_0, \mu_1, \sigma^2, x, Y, M) \quad \propto \quad \frac{exp\{-\frac{(y_i-\mu_0)^2}{2\sigma^2}\}p(g_i = 0|x, M)}{\sum_{g_i=0,1} exp\{-\frac{(y_i-\mu_{g_i})^2}{2\sigma^2}\}p(g_i|x, M)}$$

For $g_i = 1$, we have $p(g_i = 1|\mu_0, \mu_1, \sigma^2, x, Y, M) = 1 - p(g_i = 0|\mu_0, \mu_1, \sigma^2, x, Y, M)$.

Step 4: update x

The QTL location x should be sampled from $p(x|\mu_0, \mu_1, \sigma^2, G, Y, M)$. We use Metropolis-Hasting (M-H) method to do the sampling. Suppose the current location is x_k. Instead of sampling directly from the posterior distribution, M-H samples a new location x' from some other density function $Q(x'|x_k)$, and the next sample $x_{k+1} = x'$, if

$$\frac{p(x'|\mu_0, \mu_1, \sigma^2, G, Y, M))Q(x|x')}{p(x|\mu_0, \mu_1, \sigma^2, G, Y, M))Q(x'|x)} = \frac{p(x'|\mu_0, \mu_1, \sigma^2, G, Y, M))}{p(x|\mu_0, \mu_1, \sigma^2, G, Y, M))} \times \frac{Q(x|x')}{Q(x'|x)} > r, \qquad (2.21)$$

where r follows uniform distribution on [0 1], otherwise $x_{k+1} = x_k$.

In our method, we choose $Q(x'|x)$ as the uniform distribution on an interval of length 2δ around x, if it can be achieved. Thus $Q(x'|x)$ is the density function for the uniform distribution on $[max(0, x - \delta), min(x + \delta, L)]$, where L is the length of the chromosome. In the same way, Thus $Q(x|x')$ is the density function for the uniform distribution on $[max(0, x' - \delta), min(x' + \delta, L)]$. Hence,

$$\begin{aligned} Q(x'|x) &= \frac{1}{min(x + \delta, L) - max(0, x - \delta)} \\ Q(x|x') &= \frac{1}{min(x' + \delta, L) - max(0, x' - \delta)} \end{aligned} \qquad (2.22)$$

Now, we look at the ratio of conditional probabilities:

$$\begin{aligned}
& \frac{p(x'|\mu_0,\mu_1,\sigma^2,G,Y,M)}{p(x|\mu_0,\mu_1,\sigma^2,G,Y,M)} \\
= & \frac{p(\mu_0,\mu_1,\sigma^2,x',G,Y|M)}{p(\mu_0,\mu_1,\sigma^2,x,G,Y|M)} \\
= & \frac{p(Y|\mu_0,\mu_1,\sigma^2,x',G,M)p(\mu_0,\mu_1,\sigma^2,x',G,M)}{p(Y|\mu_0,\mu_1,\sigma^2,x,G,M)p(\mu_0,\mu_1,\sigma^2,x,G,M)} \\
= & \frac{p(Y|\mu_0,\mu_1,\sigma^2,G)p(G|x',M)p(x')}{p(Y|\mu_0,\mu_1,\sigma^2,G)p(G|x,M)p(x)} \qquad (2.23) \\
= & \frac{p(G|x',M)}{p(G|x',M)} \qquad (2.24) \\
= & \frac{\prod_{i=1}^n p(g_i|x',M)}{\prod_{i=1}^n p(g_i|x',M)} \qquad (2.25)
\end{aligned}$$

Equation (2.23) is because given the location x and marker information M, the distribution of G do not depend on other parameters. We obtain Equation (2.24) because the prior distribution of x ad x' is the same. Again $p(g_i|x',M)$ presents the distribution of the ith individual's genotype, given the the location as x and the marker information. Thus, given current location x_k, we get the sample x' from the uniform distribution on $[max(0,x-\delta), min(x+\delta,L)]$. Using M-H algorithm, if

$$\frac{\prod_{i=1}^n p(g_i|x',M)}{\prod_{i=1}^n p(g_i|x',M)} \times \frac{min(x+\delta,L)-max(0,x-\delta)}{min(x'+\delta,L)-max(0,x'-\delta)} > r, \qquad (2.26)$$

we set $x_{k+1}=x'$, otherwise $x_{k+1}=x_k$.

We use the sequential sampling method and get 110,000 samples for x. The first 10,000 samples are dropped (burn-in). We pick one in every 100 samples and finally get 1000 samples for x. The distribution of x is estimated using the empirical distribution of these 1000 samples.

Use a very similar way as above, we can develop similar Gibbs sampling method for mapping Single QTL of F2 population.

CHAPTER 3

Multiple QTL Model

In Chapter 3, we develop the joint multiple QTL Bayesian method based on the extension of the one QTL Bayesian method via the Laplace approximation in Chapter 2. In practice, phenotypes may be affected by several QTL not just one QTL, such as hypertension, which is a polygenic and highly variable phenotype. Thus, it is important to develop a statistical method for detecting multiple QTL. In this chapter, we will describe our joint multiple QTL Bayesian method by using the joint two QTL Bayesian model as an example. We can easily obtain the multiple QTL Model following the same idea.

There are several advantages to use the proposed Bayesian method. The first advantage of the proposed method is that we can obtain the linkage posterior probability directly, which is easy to interpret. Second, it is easy to obtain the posterior probability for all parameters using our method and get the highest posterior density (HPD) region for each parameter in the model. The HPD region is the Bayesian "confidence interval" and using HPD region, we can evaluate if the parameter is significant or not. The third advantage is that our method has the higher speed compared to the standard MCMC Method (Satagopan *et al.* (1996); Berry (1998); Sillanpaa and Arjas (1998); Stephens and Fisch (1998); Yi and S. (2000), Yi and Xu (2001), Yi (2004), Huang *et al.* (2007)) for detecting QTL. Furthermore, we develop the sequential multiple QTL Bayesian model via the Laplace approximation for quickly detecting multiple QTL locations without calculating the posterior probability. After we detect the QTL locations using this method, we can use the profile likelihood ratio test statistic to evaluate if there are multiple QTL locations or not. We use the simulation studies and real data analysis to demonstrate the proposed joint multiple QTL Bayesian method. For the sequential multiple QTL Bayesian method, a simulation study is used to

show the consistent results for detecting QTL locations for both the joint model and the sequential method.

3.1 Methods

In this section, the two QTL model is used as an example to explain our multiple QTL method. The normal linear phenotype model from Lander and Bostein (1989) is used here. Those two QTL are assumed to locate between marker intervals. Let y_i denote the phenotype value for the i^{th} individual. The linear model of the phenotype value y_i for the backcross (BC) data is described as:

$$y_i = \mu + a_1 \cdot g_i(x_1^*) + a_2 \cdot g_i(x_2^*) + \delta \cdot g_i(x_1^*) \cdot g_i(x_2^*) + \epsilon_i, \tag{3.1}$$

where a_1 is the additive effect of the first QTL, a_2 is the additive effect of the second QTL, δ is the pairwise interaction effect for both QTL. All the parameters above are unknown and will be estimated using the E-M algorithm. $g_i(x_1)$ is a numerical representation of the first QTL genotype for the i^{th} individual at position x_1 and x_1^* signifies the location of the first QTL; $g_i(x_2)$ is a numerical representation of the second QTL genotype for the i^{th} individual at position x_2 and x_2^* signifies the location of the second QTL; ϵ_i is the residual error, distributed $N(0, \sigma^2)$. We code $g_i(x_1)$ as -1 or 1 according to whether the genotype at x_1 is Aa (heterozygotic) or AA (homozygotic) and $g_i(x_2)$ as -1 or 1 according to whether the genotype at x_2 is Aa (heterozygotic) or AA (homozygotic). We use $\boldsymbol{\beta} = \{\mu, a_1, a_2, \delta, \sigma^2\}$ to represent the nuisance parameters, occupying a possibly finite region Ω for which the prior $p(\boldsymbol{\beta}) > 0$. We wish to obtain the posterior probability of the two QTL at any gene locations x_1 and x_2, given the phenotypes and the marker genotype data,

$$\begin{aligned} p(x_1, x_2|data) &= \frac{p(x_1, x_2)p(data|x_1, x_2)}{p(data)} = \frac{p(x_1)p(x_2)\int_\Omega p(data, \boldsymbol{\beta}|x_1, x_2)d\boldsymbol{\beta}}{p(data)} \\ &= \frac{p(x_1)p(x_2)\int_\Omega p(\boldsymbol{\beta})p(data|x_1, x_2, \boldsymbol{\beta})d\boldsymbol{\beta}}{p(data)}. \end{aligned} \tag{3.2}$$

Here x_1, x_2 denote the true QTL positions. So, for example, the location priors $p(x_1)p(x_2)$,

will be understood to mean $p(x_1^* = x_1)p(x_2^* = x_2)$. These priors are intentionally flexible, because in future applications it might be sensible to consider prior information from previous studies, or to place mass only on the genomic positions of genes, implicitly favoring gene-rich genomic regions. Our goal is to enable direct probability statements for the joint posterior of x_1, x_2, so that the posterior for entire regions/chromosomes may be obtained via summation or integration. Numerous Bayesian QTL methods usually use Bayes Factors to evaluate the inference, which is less formal. Nonetheless, the Bayes factor may also be easily obtained from our approach.

The right-hand side of (3.2) follows from the assumption of independence of QTL positions and effect size, $p(x_1, x_2, \boldsymbol{\beta}) = p(x_1)\ p(x_2)\ p(\boldsymbol{\beta})$. We will denote the marker positions by the vector $\mathbf{x}_\mathrm{m}$, the markers flanking x_1 by $\{x_1^{left}, x_1^{right}\}$, and the markers flanking x_2 by $\{x_2^{left}, x_2^{right}\}$. The quantity $p(data|x1, x_2, \boldsymbol{\beta})$ is the ordinary interval mapping likelihood for n individuals:

$$p(data|x_1, x_2, \boldsymbol{\beta}) = p(g(\mathrm{x_m}))\prod_{i=1}^{n}\Big[\sum_{j_1=-1,1}\sum_{j_2=-1,1} p\big(y_i|g_i(x_1) = 2j_1 - 1, g_i(x_2) = 2j_2 - 1, \boldsymbol{\beta}\big)$$
$$\times p\big(g_i(x_1) = 2j_1 - 1, g_i(x_2) = 2j_2 - 1|\boldsymbol{\beta}, x_1, x_2, g_i(x_1^{left}), g_i(x_1^{right}), g_i(x_2^{left}), g_i(x_2^{right})\big)\Big], \tag{3.3}$$

for which we use model (3.2) and Haldane's map function for genotype probabilities.

Thus far, our presentation is simply a standard Bayesian outline of the problem. In contrast to other Bayesian QTL approaches (e.g. Satagopan *et al.* (1996); Berry (1998); Sillanpaa and Arjas (1998); Stephens and Fisch (1998); Yi and S. (2000), Yi and Xu (2001), Yi (2004), Huang *et al.* (2007)), however, we state the null hypothesis in terms of the QTL positions x_1^* and x_2^*. If x_1^* and x_2^* are on the chromosomes/genome under study, the alternative hypothesis holds (i.e. H_2: $x_1^* = x_1, x_2^* = x_2$). Otherwise, the null hypothesis holds, which we denote H_0: $x_1^* = \infty, x_2^* = \infty$ and H_1: $x_1^* = x_1, x_2^* = \infty$ or $x_1^* = \infty, x_2^* = x_2$ (Doerge *et al.* (1997)). The more commonly-used form of the null hypothesis, dating at least to Lander and Bostein (1989), is a *no-gene* null specified in terms of the nuisance parameters as $\boldsymbol{\beta} \in \Omega_0 \subset \Omega$. The details of these two different null hypotheses have been

given in Chapter 2.

A second and important advantage of our null hypothesis specification is that inference for x_1 and x_2 will be relatively insensitive to the prior for $\boldsymbol{\beta}$, because $p(\boldsymbol{\beta})$ appears in both null and alternative terms in $p(data)$. In contrast, when using the no-gene null hypothesis, inference can be highly sensitive to the prior for $\boldsymbol{\beta}$, where the subspace Ω_0 is typically of lower dimension than Ω. We use a flat (proper) prior in our illustrations of the Bayesian approach, $p(\boldsymbol{\beta}) = \frac{1}{|\Omega|}$. Thus Ω must technically be finite. However, for realistic sample sizes, we can let Ω get arbitrarily large, with essentially no change in our inference.

Using the assumed prior for $\boldsymbol{\beta}$, the integral in the numerator of (3.2) becomes

$$\int_\Omega p(\boldsymbol{\beta})p(data|x_1,x_2,\boldsymbol{\beta})d\boldsymbol{\beta} = \frac{1}{|\Omega|}\int_\Omega p(data|x_1,x_2,\boldsymbol{\beta})d\boldsymbol{\beta} = \frac{1}{|\Omega|}C(x_1,x_2), \tag{3.4}$$

where $C(x_1,x_2)$ is the integrated likelihood for a fixed x_1 and x_2. The denominator of (3.2) is

$$\begin{aligned} p(data) &= \int\int_{x_1',x_2'} p(x_1')p(x_2')\{\int_\Omega p(\boldsymbol{\beta})p(data|x_1',x_2',\boldsymbol{\beta})d\boldsymbol{\beta}\}dx_1'dx_2' \\ &= \int\int_{x_1',x_2'} p(x_1')p(x_2')\frac{1}{|\Omega|}C(x_1',x_2')dx_1'dx_2', \end{aligned} \tag{3.5}$$

so the $\frac{1}{|\Omega|}$ term cancels out both in the numerator and denominator of the equation (3.2); then we get

$$p(x_1,x_2|data) = \frac{p(x_1)p(x_2)C(x_1,x_2)}{\int\int_{x_1',x_2'} p(x_1')p(x_2')C(x_1',x_2')dx_1'dx_2'} = \frac{p(x_1)p(x_2)C(x_1,x_2)}{D}, \tag{3.6}$$

where $D = \int\int_{x_1'<\infty,x_2'<\infty} p(x_1')p(x_2')C(x_1',x_2')dx_1'dx_2' + \int_{x_1'<\infty} p(x_1',\infty)C(x_1',\infty)dx_1' + \int_{x_2'<\infty} p(\infty,x_2')C(\infty,x_2')dx_2' + p(H_0)C(\infty,\infty)$, $C(x_1',x_2')$ is the integrated likelihood for nuisance parameters under fixed x_1' and x_2'; $C(x_1',\infty), C(\infty,x_2')$ are the integrated likelihoods for nuisance parameters under H_1; and $C(\infty,\infty)$ is the integrated likelihood for nuisance parameters under H_0. The denominator D is partitioned into the alternative hypothesis H_2 as well as two null hypotheses: H_1 and H_0. $p(H_0)$ is the prior for H_0.

Now it remains to get a good approximation of $C(x_1,x_2)$ for any fixed x_1, x_2, $C(\infty,x_2)$

for any fixed x_2, $C(x_1, \infty)$ for any fixed x_1 and $C(\infty, \infty)$ since they have no closed form. As we have shown in the one QTL model, the Laplace method is the best way to estimate those approximate forms.

3.1.1 The Laplace Approximation

The Laplace approximation is the method we proposed for the approximation of the integral likelihood. Using this method, the computational intensivity problem improves immeasurably. First, we want to get the approximation of $C(x_1, x_2)$ under the alternative hypothesis that both QTL reside on the chromosomes/genome under study. We start by fixing x_1, x_2 (suppressing the dependence on x_1, x_2) and defining $f(\boldsymbol{\beta}) = p(data|x_1, x_2, \boldsymbol{\beta})$. The applicability of the Laplace approximation relies on standard behavior for the log-likelihood for large sample sizes: the function is continuous, unimodal, twice differentiable, and with a maximum in the interior of Ω (Azevedo-Filho and Shachter (1994)). The Laplace approximation may be motivated by a Taylor expansion at $\hat{\boldsymbol{\beta}}$ for a fixed x_1, x_2:

$$log(f(\boldsymbol{\beta})) = log(f(\hat{\boldsymbol{\beta}})) - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\Sigma}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + O(||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^3), \tag{3.7}$$

where $\hat{\Sigma} = I^{-1}(\hat{\boldsymbol{\beta}})$ is obtained by inverting the analytically-derived information matrix at $\hat{\boldsymbol{\beta}}$ and the m.l.e. $\hat{\boldsymbol{\beta}}$ is obtained using a standard maximization routine E-M, as is routinely performed in standard interval mapping. After exponentiating each side of the above approximation and integrating over all $\boldsymbol{\beta}$, we obtain

$$C(x_1, x_2) = \int_{\boldsymbol{\beta} \in \Omega} f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \int f(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx f(\hat{\boldsymbol{\beta}})(2\pi)^{dim(\boldsymbol{\beta})/2} |\hat{\Sigma}|^{1/2} \equiv \hat{C}(x_1, x_2). \tag{3.8}$$

The indefinite integral assumes the space Ω is "large," and the constants $(2\pi)^{dim(\boldsymbol{\beta})/2}|\hat{\Sigma}|^{1/2}$ on the fourth term in equation (3.8) arises from the integration over a multivariate normal density with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\Sigma}$. The value $f(\hat{\boldsymbol{\beta}})$ is simply the likelihood at (x_1, x_2, $\hat{\boldsymbol{\beta}}$), which is already available after deriving the standard E-M estimation of the nuisance parameters $\hat{\boldsymbol{\beta}}$ at given location x_1, x_2. We estimate $\hat{\Sigma}$ by plugging $\hat{\boldsymbol{\beta}}$ (and the known recombination fractions to the nearby markers) into the analytically-derived observed information matrix. Finally, we substitute $\hat{C}(x_1, x_2)$ for $C(x_1, x_2)$ in equation

(3.6) for $x_1 < \infty, x_2 < \infty$.

3.1.2 Approximating the null integrated likelihood

The Laplace approximation is used to estimate the null values $C(\infty,\infty), C(\infty,x_2), C(x_1,\infty)$ in this section. In Chapter 2, we have already shown that the one QTL null likelihood can not obtain an accurate estimation by applying the Laplace approximation directly. The H_0 likelihood for BC data is a mixture of four normal densities, with each of four genotype probabilities in equation (3.3) replaced by 1/4. In addition to the curvature in the likelihood contours, the likelihood can remain relatively high and flat spanning $a_1 = 0, a_2 = 0, \delta = 0$, and it is difficult to prescribe a parameter transformation that will make the likelihood approximately normal in shape. Furthermore, if such a transformation were available, it would be non-linear, and difficult to transform back to integration over the original Ω.

We use the idea of the *improved null* Laplace approximation in Chapter 2 to estimate the integration of the H_0 likelihood $C(\infty,\infty)$ for the joint Bayesian multiple QTL method. We devise the following approximations requiring integration over three parameters, using the fact that the Laplace approximation for $\{\mu,\sigma^2\}$ works well for fixed a_1, a_2 and δ. Define $f(a_1,a_2,\delta,\mu,\sigma^2) = p(\text{data}|x_1=\infty, x_2=\infty, \mu, \sigma^2, a_1, a_2, \delta)$, and $\hat{\mu}_{a_1,a_2,\delta}, \hat{\sigma}^2_{a_1,a_2,\delta}$ (obtained numerically) as the conditional m.l.e.s for fixed a_1, a_2, δ, with corresponding covariance matrix estimate $\hat{\Sigma}_{a_1,a_2,\delta}$ on the restricted space. We then have the *improved null* Laplace approximation of $\hat{C}(\infty,\infty)$ written as:

$$\begin{aligned}\hat{C}(\infty,\infty) &= \int\int\int_{a_1,a_2,\delta}[\int\int_{\mu,\sigma^2} f(a_1,a_2,\delta;\mu,\sigma^2)d\mu d\sigma^2]d(a_1)d(a_2)d(\delta)\\ &= \int\int\int_{a_1,a_2,\delta} f(a_1,a_2,\delta,\hat{\mu}_{a_1,a_2,\delta},\hat{\sigma}^2_{a_1,a_2,\delta})2\pi|\hat{\Sigma}_{a_1,a_2,\delta}|^{1/2}d(a_1)d(a_2)d(\delta).\end{aligned}$$

The H_1 likelihood for BC data is a mixture of four normal densities, with each of four genotype probabilities in equation (3.3) replaced by $1/2\times p\big(g_i(x)=k|\boldsymbol{\beta}, g_i(x^{left}), g_i(x^{right})\big)$ depending on x equal to x_1 or x_2, where k is the genotype of the QTL on the chromosomes under study and x_{left}, x_{right} are the flanking markers on the left and right for QTL.

Similarly we use the idea of the *improved null* Laplace approximation to estimate the integration of the H_1 likelihood $C(x1,\infty)$ and $C(\infty,x2)$. For $C(x1,\infty)$, we devise the ap-

proximation requiring integration over three parameters, using the fact that the Laplace approximation for $\{\mu, \sigma^2\}$ works well for fixed a_1, a_2 and δ. Define $f(a_1, a_2, \delta, \mu, \sigma^2) = p(\text{data}|x_1, x_2 = \infty, a_1, a_2, \delta, \mu, \sigma^2)$, and $\hat{\mu}_{a_1,a_2,\delta}, \hat{\sigma}^2_{a_1,a_2,\delta}$ (obtained numerically) as the conditional m.l.e.s for fixed a_1, a_2, δ, with a corresponding covariance matrix estimate $\hat{\Sigma}_{a_1,a_2,\delta}$ on the restricted space. We then have the *improved null* Laplace approximation of $\hat{C}(x_1, \infty)$ written as:

$$\begin{aligned}\hat{C}(x_1, \infty) &= \int\int\int_{a_1,a_2,\delta} [\int\int_{\mu,\sigma^2} f(a_1, a_2, \delta, \mu, \sigma^2) d\mu d\sigma^2] d(a_1) d(a_2) d(\delta) \\ &= \int\int\int_{a_1,a_2,\delta} f(a_1, a_2, \delta, \hat{\mu}_{a_1,a_2,\delta}, \hat{\sigma}^2_{a_1,a_2,\delta}) 2\pi |\hat{\Sigma}_{a_1,a_2,\delta}|^{1/2} d(a_1) d(a_2) d(\delta).\end{aligned}$$

In the same way, we can obtain $\hat{C}(\infty, x_2)$.

3.1.3 Posterior Curves for All Nuisance Parameters

The advantage of using the Bayesian method is that we can calculate the posterior densities for all parameters β as well as their highest posterior density (HPD) regions. The following is the formula of $p(\boldsymbol{\beta}|data)$:

$$p(\boldsymbol{\beta}|data) = \int_{all x_1, x_2} p(\boldsymbol{\beta}|x_1, x_2, data) p(x_1, x_2|data) d(x_1) d(x_2). \tag{3.9}$$

We already use the laplace approximation to estimate $p(x_1, x_2|data)$. The posterior density of β given x_1, x_2 can be approximated by

$$\hat{p}(\hat{\boldsymbol{\beta}}|x_1, x_2, data) = \frac{1}{(2\pi)^{d/2} |\hat{\Sigma}|^{1/2}} exp(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\hat{\Sigma})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})),$$

where $\hat{\Sigma} = I^{-1}(\hat{\boldsymbol{\beta}})$ is obtained by inverting the analytically-derived information matrix at $\hat{\boldsymbol{\beta}}$ and the m.l.e. $\hat{\boldsymbol{\beta}}$ is obtained using a standard maximization routine E-M; d represents the number of the nuisance parameters $\boldsymbol{\beta}$. This approximation uses the Taylor expansion of $p(\boldsymbol{\beta}|x_1, x_2, data)$ around $\hat{\boldsymbol{\beta}}$ for each x_1, x_2. The approximated density is multivariate Gaussian distribution.

We obtain $p(\boldsymbol{\beta}|data)$ from the equation (3.9). The joint posterior probability $p(\boldsymbol{\beta}|data)$ is the joint normal distribution, so the posterior probability of each parameter is normally distributed, too. Because we know the posterior probability for each parameter is approximately normal, which is symmetric, we can calculate the 95% HPD region for each parameter using the following formula: the parameter posterior mean $\pm$ 1.96$*\sqrt{\textit{the parameter posterior variance}}$. If the 95% HPD region for each parameter does not include 0, this means the parameter has the effect for the model.

3.1.4 Sequential Multiple QTL Bayesian Model

In order to save computation time for quickly finding multiple QTL locations without calculating the posterior probability, we developed the sequential multiple QTL heuristic to overcome this problem. We use the two QTL model to illustrate our idea of this method. The algorithm is as follows:

(1) First, we find the first estimated QTL location on the chromosomes under study by using our one QTL Bayesian method via the Laplace approximation.

(2) Second, conditional on the first QTL location we detected, we use the two QTL Bayesian method without considering the null hypothesis (the partial two QTL Bayesian method) to find the second QTL location on the chromosomes under study.

(3) Then conditional on the second QTL location we detected, we use the partial two QTL Bayesian method again to find the first QTL location on the chromosomes under study.

Then iteratively repeat the process above until the QTL estimated locations converge. We finally can find two QTL locations quickly, instead of calculating the joint posterior probability of two QTL.

Figure 3.1 provides an example to illustrate our method. We study two chromosomes for detecting two QTL locations. First, we detect the first QTL location using the one QTL Bayesian method via the Laplace approximation. We find that the location of the first QTL is at 10.5cM on the first chromosome. Second, given that the first QTL location is at 10.5cM on the first chromosome, we use the partial two QTL Bayesian model to detect the second QTL location, which is at 80.5cM on the second chromosome. Then conditional

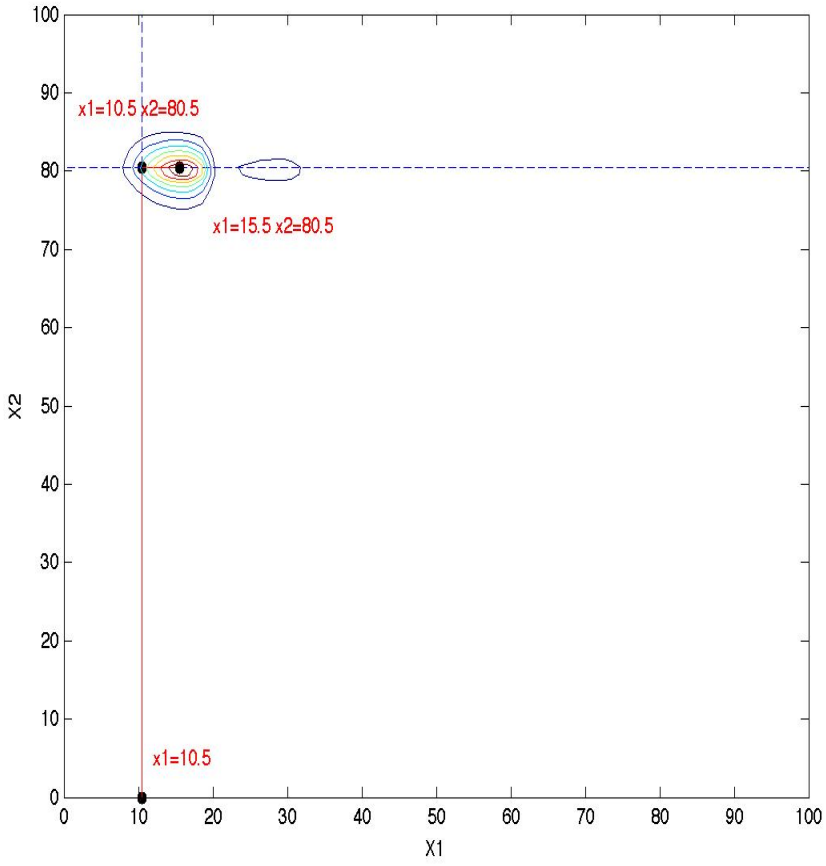

Figure 3.1: The sequential Bayesian QTL method algorithm for detecting two QTL.

on this second QTL location at 80.5cM on the second chromosome, we go back to look for the location of the first QTL, which is at 15.5cM on the first chromosome. Finally, we condition on this first QTL location and find the second QTL, which is still at 80.5cM on the second chromosome. This means that our process has converged to the first QTL location at 15.5cM on the first chromosome and the second QTL location at 15.5cM on the second chromosome. In this way, we can obtain the QTL locations quickly instead of calculating the joint posterior probability for all putative QTL locations. In order to test if the QTL locations we detect are significant or not, we can use the log_{10} of the likelihood ratio test statistic.

3.2 Simulation Studies

We use the simulation studies to demonstrate the proposed joint Bayesian multiple QTL model and the proposed sequential Bayesian multiple QTL model.

3.2.1 Simulation Results for the Joint Bayesian Multiple QTL Model

We generate 100 BC data sets from each of the hypotheses: H_0, H_1 and H_2 respectively to evaluate our joint Bayesian two QTL method. In Figure 3.2, nine histograms show the simulation results. Histograms in the first row show the posterior simulation results for 100 data sets generated from H_0. Histograms in the second row are the posterior simulation results for 100 data sets generated from H_1. Histograms in the third row are the posterior simulation results for 100 data sets generated from H_2. The most left plot on the first row shows the histogram of 100 H_0 posterior probabilities, i.e. $p(H_0|data)$, obtained using our method from 100 H_0 data sets. The middle plot on the first row shows the histogram of 100 H_1 posterior probabilities, i.e. $p(H_1|data)$, obtained by our method from 100 H_0 data sets. The most right plot on the first row shows the histogram of 100 H_2 posterior probabilities, i.e. $p(H_2|data)$, obtained from 100 H_0 data sets by using our method. Because the 100 data sets are all generated from H_0, we can see that $p(H_0|data)$ for all data sets have higher posterior probabilities compared to $p(H_1|data)$ and $p(H_2|data)$ histogram plots by using our method. We can also use the same procedure to interpret the histograms on the second row, where the data are generated from H_1. The far left plot on the second row shows

the histogram of 100 H_0 posterior probabilities, i.e. $p(H_0|data)$, obtained by our method from 100 H_1 data sets. The middle plot on the second row shows the histogram of 100 H_1 posterior probabilities, i.e. $p(H_1|data)$, obtained by our method from 100 H_1 data sets. The far right plot on the second row shows the histogram of 100 H_2 posterior probabilities, i.e. $p(H_2|data)$, obtained by our method from 100 H_1 data sets. Because the 100 data sets are all generated from H_1, the $p(H_1|data)$ for all data sets have higher posterior probabilities compared to $p(H_0|data)$ and $p(H_2|data)$ plots by using our method. For the histograms on the third row, we can use the same procedure to explain them. Because the 100 data sets are all generated from H_2 on the third row, the $p(H_2|data)$ for all data sets have higher posterior probabilities compared to $p(H_0|data)$ and $p(H_1|data)$ plots by using our method. The simulation results for all nine histograms demonstrate that our method works well for multiple QTL analysis.

A Receiver Operating Characteristic (ROC) curve is a plot to show the true positive rate against the false positive rate for different possible cut points of a diagnostic test. This curve is for binary outcomes. The area under the ROC curve (AUC) is a measure to evaluate how accurate the method is. If the measure AUC is nearly 1, this means the method is an excellent test. But if AUC is around or below 0.5, this means that the method is a worthless test. In Figure 3.3, we use the ROC curve to evaluate our methods for detecting posterior probability $p(H_0|data)$, $p(H_1|data)$ and $p(H_2|data)$ under different thresholds. The left plot on the first row is to evaluate $p(H_0|data)$ vs. $p(H_1|data)$ plus $p(H_2|data)$ for 100 simulated data generated from H_0. AUC in this plot is nearly 1, therefore our method is a excellent test for detecting $p(H_0|data)$. The right plot on the first row is to evaluate $p(H_1|data)$ vs. $p(H_0|data)$ plus $p(H_2|data)$ for 100 simulated data generated from H_1. The AUC in this plot is also nearly 1, thus our method for detecting $p(H_1|data)$ is also very accurate. Similarly, the plot on the second row is to evaluate $p(H_2|data)$ vs. $p(H_1|data)$ plus $p(H_0|data)$ for 100 simulated data generated from H_2. AUC measure in this plot is again nearly 1 so it means our method is a good test for detecting $p(H_2|data)$.

We also simulate 100 BC data sets under the following condition that the first QTL location is at 25 cM, the second QTL location is at 75 cM, the first QTL effect a_1 is 0.5, the second QTL effect a_2 is 0.5 and their interaction effect δ is 0.1. We use our

joint Bayesian multiple QTL method to evaluate whether our method can detect the QTL locations accurately. In Figure 3.4, the x axis is the estimated location for the first QTL, the y axis is the estimated location for the second QTL, and there are 100 points on the plot, which shows the estimated QTL locations for 100 simulated data sets using our joint Bayesian multiple QTL method. The average estimated first QTL location from 100 data sets is 27.71 cM and the average estimated second QTL location from 100 data sets is 74.10 cM. The average estimated QTL values are very close to the true QTL locations, so we can conclude that our method detecte the QTL locations accurately.

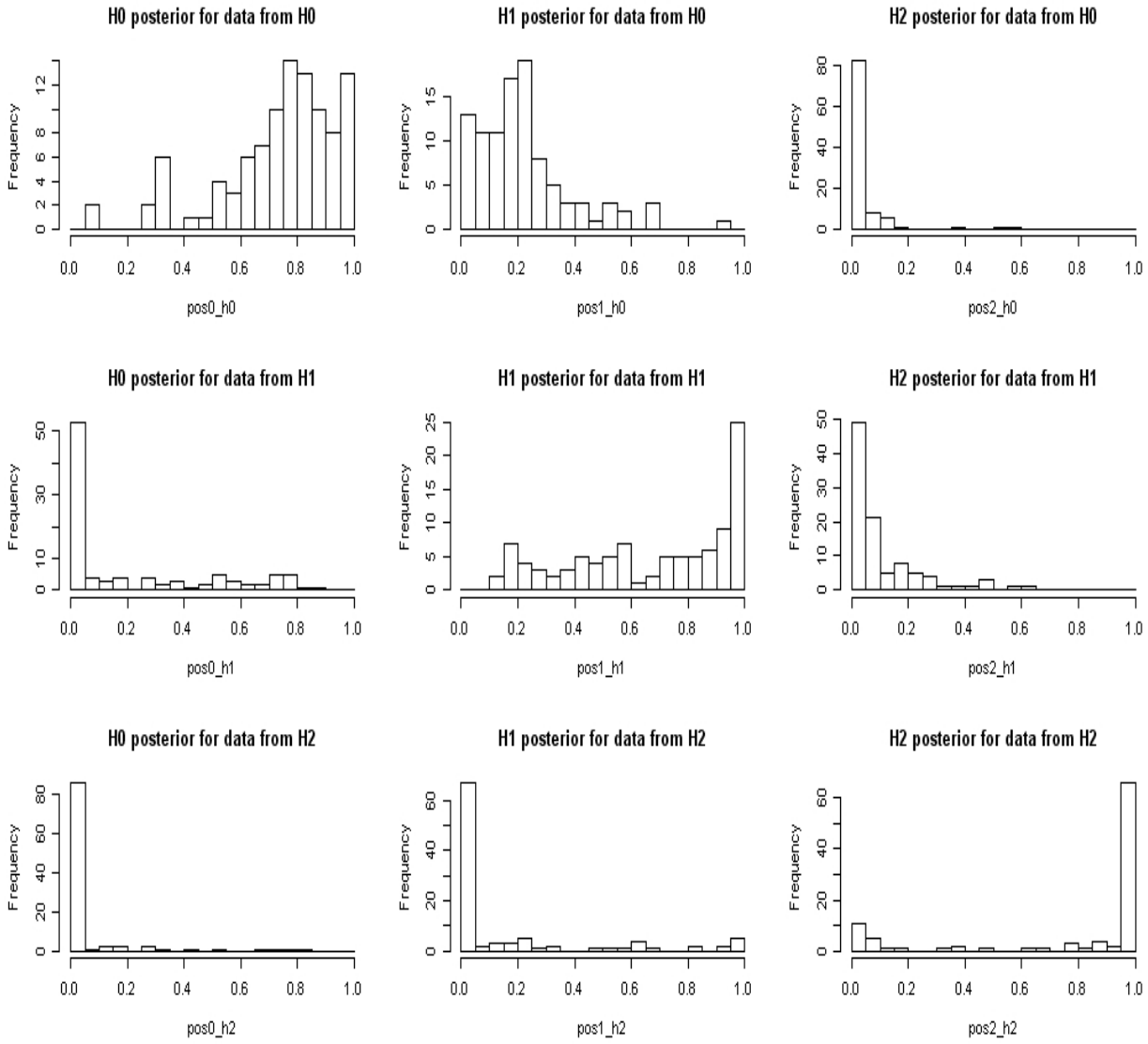

Figure 3.2: 300 simulation results for data generated from H_0, H_1 and H_2.

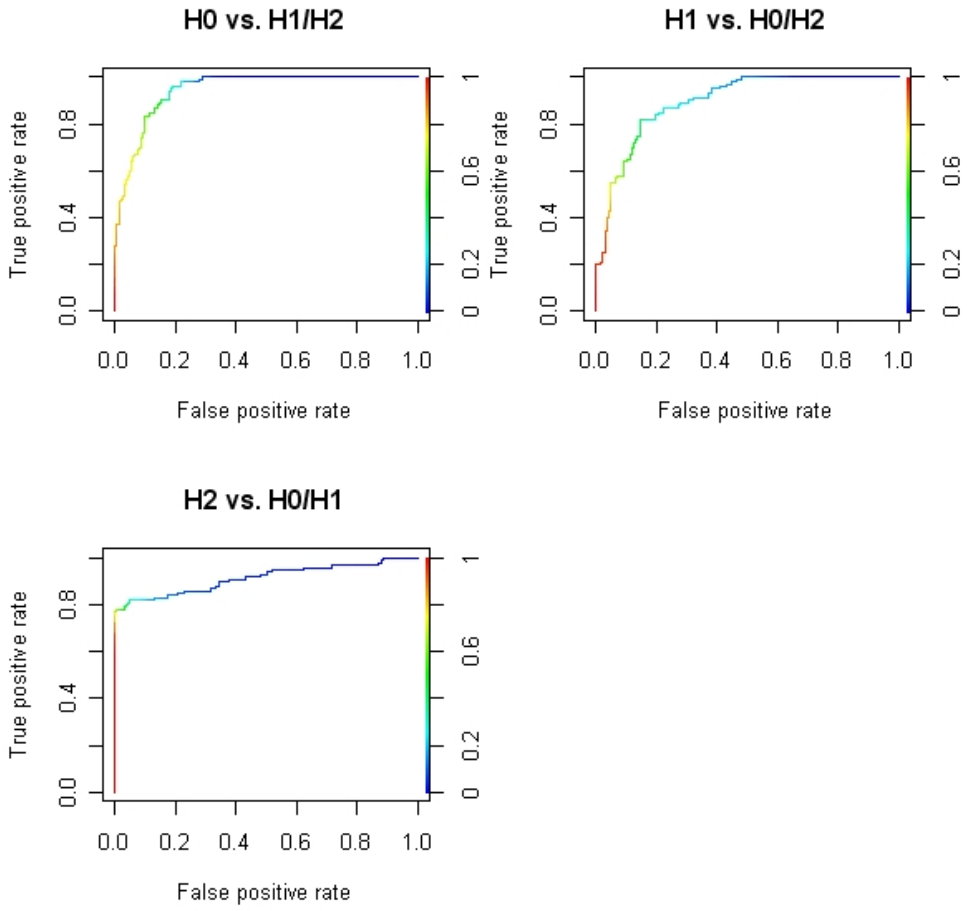

Figure 3.3: ROC curve for all three hypotheses.

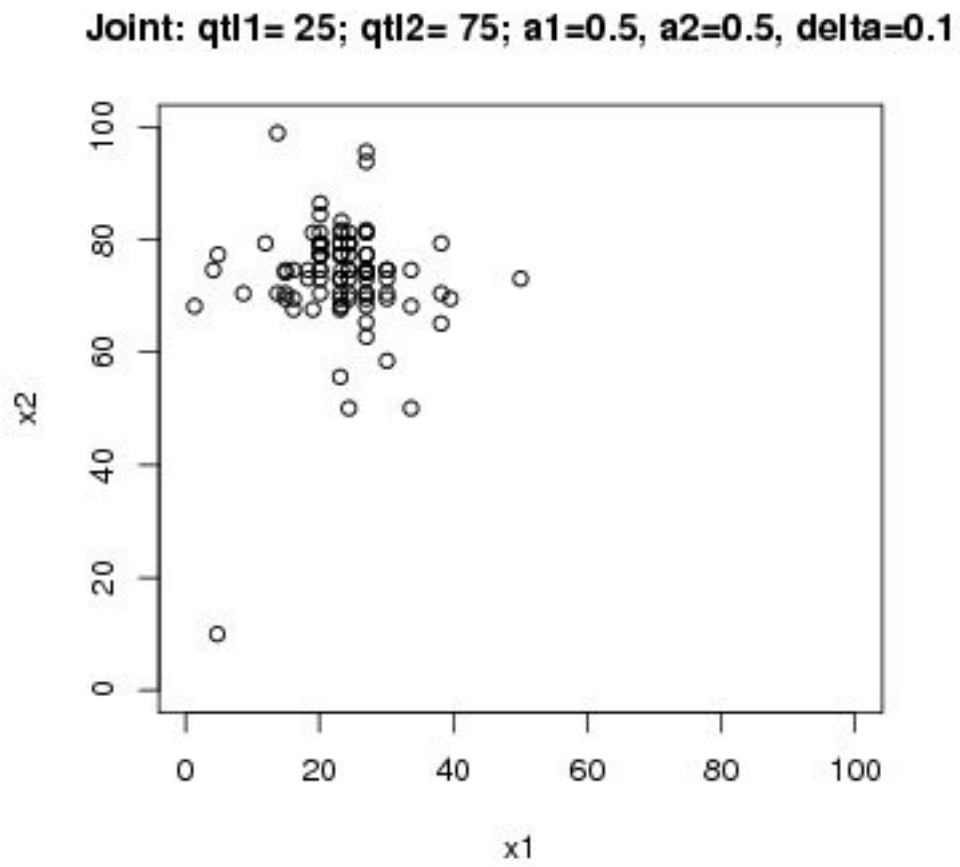

Figure 3.4: 100 simulation results for detecting QTL locations by using joint Bayesian two QTL method.

3.2.2 Simulation Results for Sequential Bayesian Multiple QTL model

We developed the sequential Bayesian multiple QTL method mainly to look for QTL locations quickly, so it is important to know whether this method can detect the QTL locations accurately. We use the same 100 simulated data sets (which we used before to evaluate how accurately the joint Bayesian multiple QTL method could detect the true QTL) to evaluate the proposed sequential method and show the simulated result in Figure 3.5. In Figure 3.5, the x axis is the estimated location of the first QTL, the y axis is the estimated location of the second QTL, and there are 100 points on the plot, which displays the estimated QTL locations for 100 simulated data sets, by using the proposed sequential Bayesian multiple QTL method. The average estimated first QTL location from the 100 data sets is 26.47 cM and the average estimated second QTL location from the 100 data sets is 74.38 cM. The average estimated QTL values are very close to the true QTL locations, so we conclude that the sequential method can detect the QTL locations accurately.

We also compare the simulated results of both the joint and sequential methods and are interested in that if they can detect consistent estimated QTL locations. Figure 3.6 shows the comparison results for both methods. The upper plot shows the simulated results of the first QTL locations for 100 data sets. The x axis is the estimated first QTL location obtained from the joint method and the y axis is the estimated first QTL location obtained from the sequential method. The lower plot shows the simulated results of the second QTL locations for 100 data sets. The x axis is the estimated second QTL location obtained from the joint method and the y axis is the estimated second QTL location obtained from the sequential method. Most of the estimated first QTL locations are consistent for both methods except that one simulation result has some variation. For the simulation results of the second QTL locations, both methods have quite similar estimated locations. From Figures 3.3 to 3.5, we therefore can conclude that both methods have high accuracy and give consistent results for detecting QTL locations. However, the speed of the sequential Bayesian multiple QTL method is much higher than the speed of the joint Bayesian multiple QTL method.

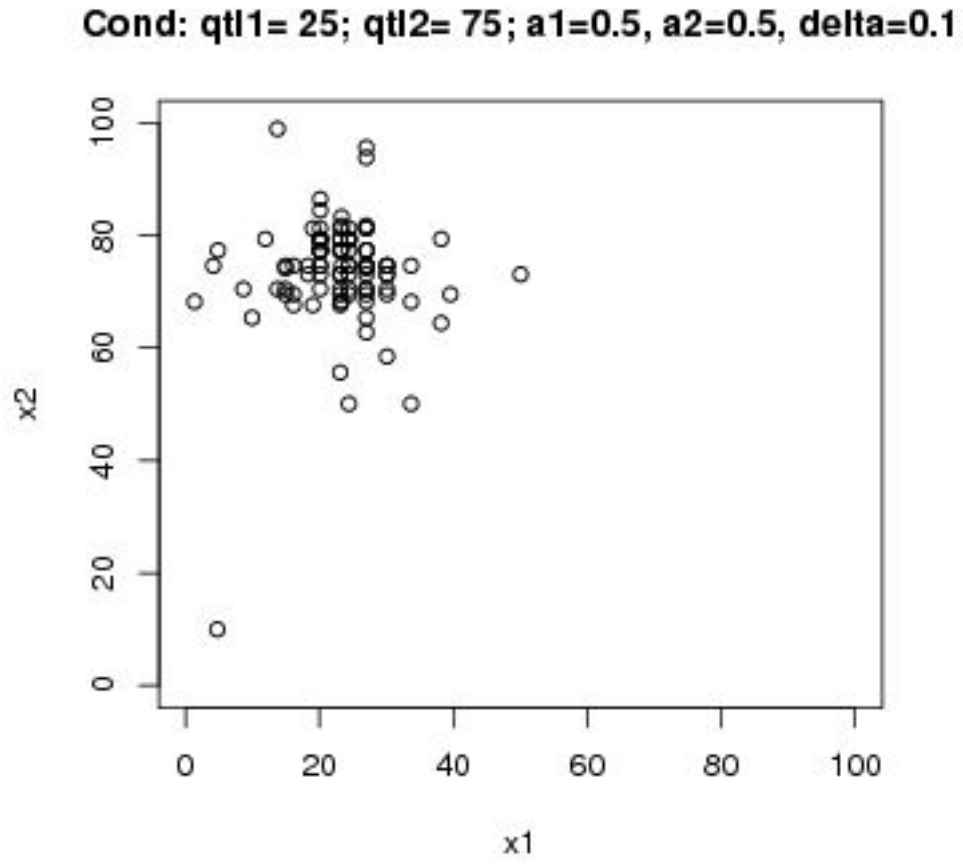

Figure 3.5: 100 simulation results for detecting QTL locations by using sequential Bayesian two QTL method.

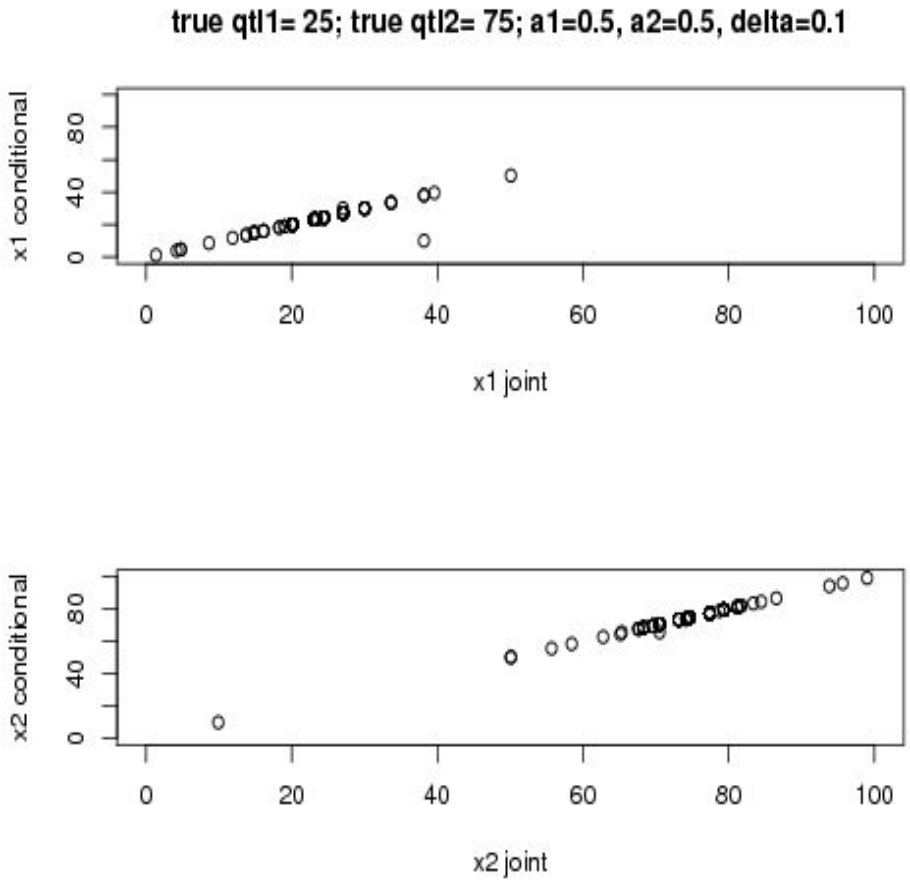

Figure 3.6: Compare 100 simulation results of joint method and sequential method for detecting QTL locations.

3.3 Real Data Analysis

We apply the proposed joint Bayesian multiple QTL model to the BC data from Sugiyama *et al.* (2001). The phenotype in this paper is the salt-induced hypertension - blood pressure measurement. Hypertension is a polygenic, complicated and highly variable trait, so this phenotype is suitable for our joint Bayesian multiple QTL method. This BC data set is the male progeny obtained from a cross of salt-sensitive C57BL/6J (B6) and non-salt-sensitive A/J (A) inbred mouse strains. These two mouse strains produced 250 male BC progeny. Females are not included in our analysis because salt increases blood pressure more in B6 males than in B6 females. From the results of Sugiyama *et al.* (2001), there are some significant QTL linked to the salt-induced hypertension phenotype. On chromosome 6, there is one significant QTL with a main effect on the phenotype; on chromosome 15, there is also one significant QTL with a main effect on the phenotype, and both QTL on chromosomes 6 and 15 have an interaction effect. We conducted the real data analysis on these two chromosomes, because our methods can deal with the situation where there are multiple QTL with interaction effects on the chromosomes under study.

Chromosome 6 is 80 cM long with 11 markers and chromosome 15 is 70 cM long with 11 markers. The 250 BC mice are genotyped but many of them have the missing marker information. If the genotype information for some markers is missing, the nearest non-missing genotype marker can be used as an alternative flanking marker. The phenotypic value is standardized in our data analysis. We generated 100 putative QTL locations uniformly across chromosomes 6 and 11 respectively to evaluate the QTL posterior probabilities at each location. We chose $\frac{1}{3}$ as the prior for H_0 and $\frac{1}{600}$ for each putative location under H_1. The prior for the QTL location is specified as $\frac{1}{30000}$ for H_2.

In Figure 3.7, we plot the posterior probabilities for 10000 putative locations on chromosomes 6 and 15. The x axis is the putative QTL location for chromosome 6, the y axis is the putative QTL location for chromosome 15, and the z axis shows the posterior probability. We found that the first estimated QTL location is at 73.76 cM on chromosome 6, which is very close to the significant marker D6Mit15 at 74cM on chromosome 6 in the paper Sugiyama *et al.* (2001); the second estimated QTL location is at 19.41 cM on chromosome

15, which is also very close to the significant marker D15Mit152 at 20.2 cM on chromosome 15 in the paper Sugiyama *et al.* (2001). The posterior probability of $H2$ (i.e. $p(H_2|data)$) is 0.9751, which is very close to 1. The posterior probability of $H1$ (i.e. $p(H_1|data)$) is 0.0187 and the posterior probability of $H0$ (i.e. $p(H_0|data)$) is 0.0062. Both H_0 and H_1 posterior probabilities are very small. Therefore, there is strong evidence that there are two QTL on chromosomes 6 and 15 based on the information that $p(H_2|data)$ is almost 1. In Figure 3.8, we use a contour plot to show our real data analysis results. In this contour plot, the x axis is the putative QTL location for chromosome 6, and the y axis is the putative QTL location for chromosome 15. We can see very clearly that there is a peak at the QTL location 73.76 cM on chromosome 6 and at 19.41 cM on chromosome 15.

In Figure 3.9, we draw the posterior probability curve for all nuisance parameters μ, a_1, a_2, δ and σ^2. From the posterior probability curve of μ, we obtain the posterior mean of the parameter μ is 0.0403 and the posterior variance is 0.0037. The posterior mean of the parameter a_1 is 0.1995 and the posterior variance is 0.0046, so we know the additive effect of the first QTL is positive with a value 0.1995 on chromosome 6. The posterior mean of the parameter a_2 is -0.2037 and its posterior variance is 0.0042; this means that the second QTL has a negative effect around -0.2 on chromosome 15. The posterior mean of the parameter δ has a negative effect around -0.2875 and the posterior variance is 0.0045, thus these two QTL have a negative interaction effect with a value -0.2875. The interaction effect is consistent with the results in paper Sugiyama *et al.* (2001): the two QTL have an interaction effect which is negative. The posterior mean of the parameter σ^2 is 0.8180 with a posterior variance of 0.0309.

From the information above, we can calculate the highest posterior density (HPD) intervals for all parameters. The HPD interval for μ is (-0.0783, 0.1589), which includes 0, so the parameter μ has no effect for QTL. The HPD interval for a_1 is (0.0665, 0.3324), which does not include 0, so the model has a positive effect for the first QTL. The HPD interval for a_2 is (-0.3315, -0.0759), which does not include 0, so the model has a negative effect for the second QTL. The HPD interval for δ is (-0.4317, -0.1433), which does not include 0, so the model has a negative interaction effect for these two QTL. The HPD interval for σ^2 is (0.4732, 1.1628), which does not include 0, so the model shows that there is significant

environmental variation.

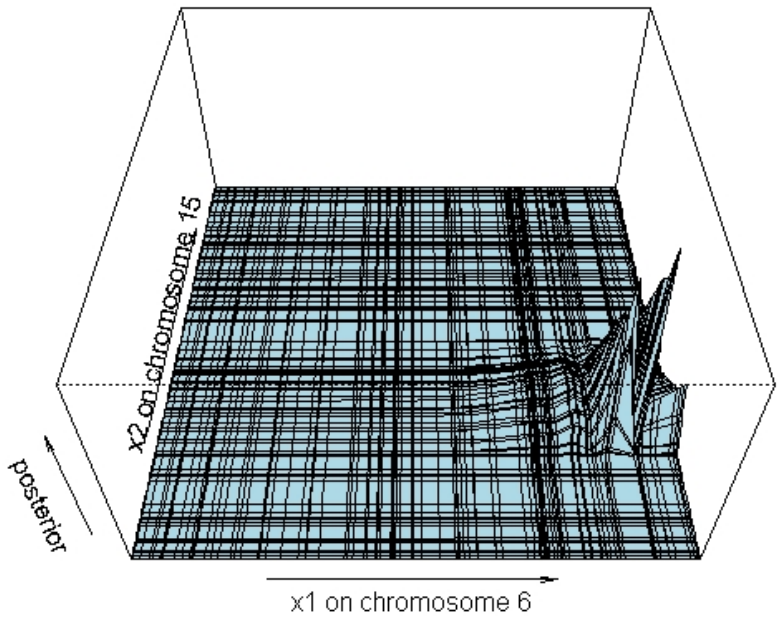

Figure 3.7: Posterior distributions of QTL locations on the chromosome 6 and 15 of the BC data from paper Sugiyama *et al.* (2001). Joint two QTL Bayesian method is used in this real data analysis.

3.4 Conclusions

We proposed the joint Bayesian multiple QTL method to detect QTL that affect the phenotype in which we are interested . By using our method, we not only can find the significant QTL that affect the phenotype faster than the standard MCMC method but also can obtain the posterior probability for inference. It is easy to interpret the statistical results using the posterior probability directly. Most Bayesian methods use the Bayes factor to interpret their results, which is less formal. Using our method, we also can get the Bayes factor. In the simulation results and real data analysis, our proposed joint method can

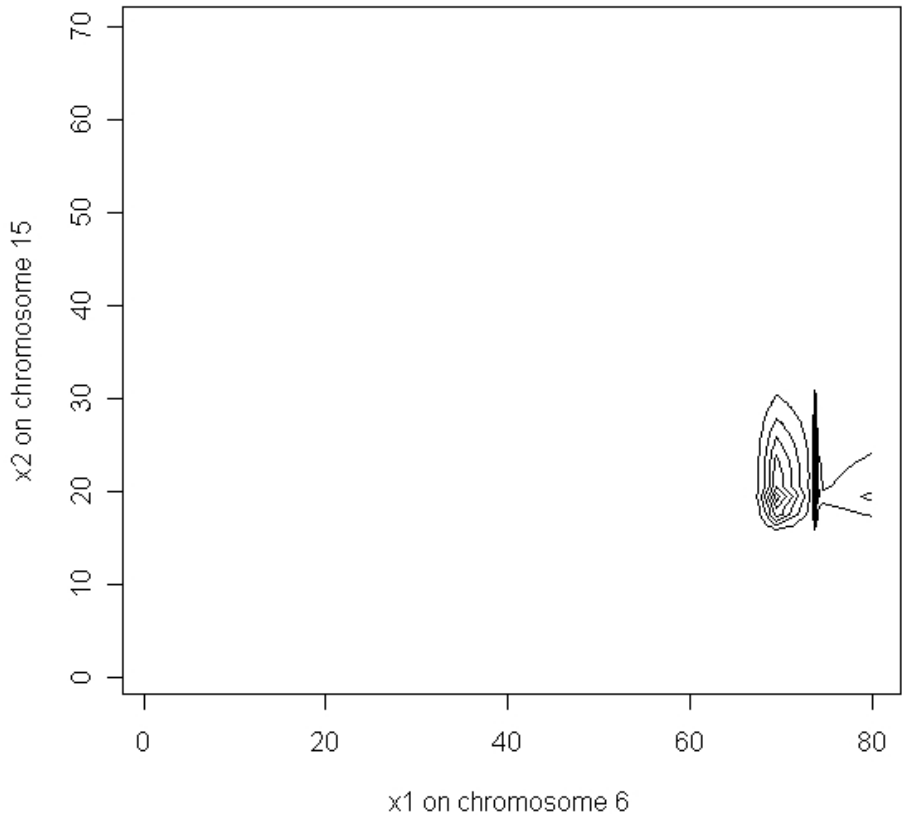

Figure 3.8: The contour plot of QTL locations on the chromosome 6 and 15 of the BC data from paper Sugiyama *et al.* (2001). Joint two QTL Bayesian method is used in this real data analysis.

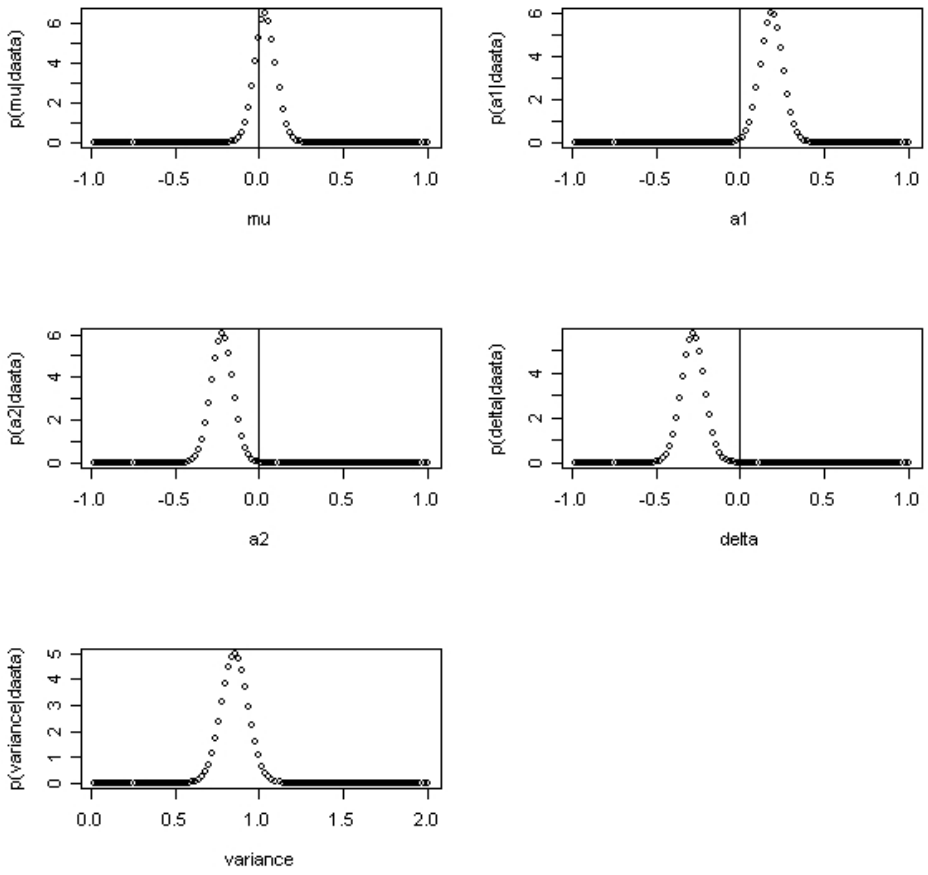

Figure 3.9: The posterior probability curve of all the parameters in real data analysis.

detect QTL locations accurately and obtain the correct linkage posterior probability for the hypothesis. Therefore, we can conclude that our method has great accuracy and moderate speed for multiple QTL analysis. In future work, we may extend the idea of this method in eQTL analysis to the detection of multiple eQTL and look for a faster null hypothesis approximation for our joint Bayesian multiple QTL method.

We also developed a sequential Bayesian multiple QTL method for detecting QTL locations that affect the phenotype we are interested in without calculating the posterior probability. In the simulation study, this method results that are consistent with the joint Bayesian multiple QTL method for detecting significant QTL locations. This means that it has high accuracy for detecting QTL locations and the method only takes several seconds to find the QTL locations. In contrast, the joint Bayesian method can take several hours, depending on how complicated the QTL model is. Therefore, we conclude that by using this sequential Bayesian multiple QTL method, we can save much computation time over other methods and quickly and accurately find QTL locations.

Extensions for full multiple QTL Bayesian methods to the high-throughput setting are a high-priority, and we plan extensions based on our conditional-search heuristic.

3.5 Appendix: Fisher Information matrix under H_A for Backcross in Two QTL Model

The alternative hypothesis in this section is that the locations of 2 QTL are on the chromosome under study and nuisance parameters are $\boldsymbol{\beta} = \{\mu,\ a_1,\ a_2,\ \delta\}$. Likelihood function for the i_{th} individual under the alternative hypothesis can be expressed as:

$$\begin{aligned} f(y_i|\boldsymbol{\beta}, x) &= \sum_{j_1=0}^{1}\sum_{j_2=0}^{1} f(y_i|k_1 = 2j_1 - 1, k_2 = 2j_2 - 1, \boldsymbol{\beta}, x_1, x_2) \\ &\times\ p(k_1 = 2j_1 - 1, k_2 = 2j_2 - 1|\boldsymbol{\beta}, x_1, x_2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\{exp(-\frac{(y_i - \mu + a_1 + a_2 - \delta)^2}{2\sigma^2})g_0(x_1, x_2) \\ &+\ exp(-\frac{(y_i - \mu + a_1 - a_2 + \delta)^2}{2\sigma^2})g_1(x_1, x_2) \\ &+\ exp(-\frac{(y_i - \mu - a_1 + a_2 + \delta)^2}{2\sigma^2})g_2(x_1, x_2) \\ &+\ exp(-\frac{(y_i - \mu + a_1 + a_2 - \delta)^2}{2\sigma^2})g_3(x_1, x_2)\}, \end{aligned}$$

where

$$g_0(x) = p(k_1 = -1, k_2 = -1|\boldsymbol{\beta}, x_1, x_2),\ g_1(x) = p(k_1 = -1, k_2 = 1|\boldsymbol{\beta}, x_1, x_2),$$

$$g_2(x) = p(k_1 = 1, k_2 = -1|\boldsymbol{\beta}, x_1, x_2),\ g_3(x) = p(k_1 = 1, k_2 = 1|\boldsymbol{\beta}, x_1, x_2),$$

For simplicity, we reparameterize the nuisance parameters from $\boldsymbol{\beta} = \{\mu,\ a_1,\ a_2,\ \delta\}$ to $\hat{\boldsymbol{\beta}} = \{\mu 0,\ , \mu 1,\ , \mu 2\ , \mu 3\}$ and those μs represent the means of different populations, which depend on two QTL genotypes.

Likelihood function can be further expressed as

$$
\begin{aligned}
f(y_i|\boldsymbol{\beta}, x) = f(y_i|\hat{\boldsymbol{\beta}}, x) &= \sum_{j_1=0}^{1}\sum_{j_2=0}^{1} f(y_i|k_1 = 2j_1 - 1, k_2 = 2j_2 - 1, \hat{\boldsymbol{\beta}}, x_1, x_2) \\
&\times \; p(k_1 = 2j_1 - 1, k_2 = 2j_2 - 1|\hat{\boldsymbol{\beta}}, x_1, x_2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\{exp(-\frac{(y_i - \mu 0)^2}{2\sigma^2})g_0(x_1, x_2) + exp(-\frac{(y_i - \mu 1)^2}{2\sigma^2})g_1(x_1, x_2) \\
&+ \; exp(-\frac{(y_i - \mu 2)^2}{2\sigma^2})g_2(x_1, x_2) + exp(-\frac{(y_i - \mu 3)^2}{2\sigma^2})g_3(x_1, x_2)\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\mathbf{A}_i,
\end{aligned}
$$

where

$$\mu 0 = \mu - a_1 - a_2 + \delta, \;\; \mu 1 = \mu - a_1 + a_2 - \delta,$$

$$\mu 2 = \mu + a_1 - a_2 - \delta, \;\; \mu 4 = \mu + a_1 + a_2 + \delta,$$

$$\mathbf{A}_i = \sum_{s=0}^{3} exp(-\frac{(y_i - \mu_s)^2}{2\sigma^2})g_s(x), \;\; s = 0, 1, 2, 3.$$

The likelihood for all the individuals are :

$$f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, x) = \prod_{i=1}^{n} f(y_i|\hat{\boldsymbol{\beta}}, x) = (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \mathbf{A}_i$$

After taking log for the likelihood above:

$$\ell = \ln f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, x) \propto \frac{-n}{2} \ln \sigma^2 + \sum_{i=1}^{n} \ln \mathbf{A}_i$$

Therefore,

$$\frac{\partial log(f(\boldsymbol{y}|\boldsymbol{\beta}, x))}{\partial \beta} = \frac{\partial \ell}{\partial \hat{\beta}} \frac{\partial \hat{\beta}}{\partial \beta}$$

$$\Rrightarrow |\frac{\partial^2 log(f(\boldsymbol{y}|\boldsymbol{\beta}, x))}{\partial \beta^2}|_{mle} = |\frac{\partial^2 \ell}{\partial \hat{\beta}^2}|_{mle}|J(\hat{\beta}|\beta)|^2.$$

In the following sections, $\frac{\partial^2 \ell}{\partial \hat{\beta}^2}$ is derived and $|J(\hat{\beta}|\beta)|$ =16.

3.5.1 The First Derivatives of the Loglikelihood Function

We define

$$\mathbf{B}_i(s) \doteq \frac{\partial \mathbf{A}_i}{\partial \mu_s} = g_s(x) exp(-\frac{(y_i-\mu_s)^2}{2\sigma^2})\frac{(y_i-\mu_s)}{\sigma^2},$$

where $s = 0, 1, 2, 3$.

and

$$\begin{aligned}
\mathbf{C}_i &\doteq \frac{\partial \mathbf{A}_i}{\partial \sigma^2} \\
&= \sum_{s=0}^{3} g_s(x) exp(-\frac{(y_i-\mu_s)^2}{2\sigma^2})\frac{(y_i-\mu_s)^2}{2(\sigma^2)^2} \\
&= \frac{1}{2\sigma^2}\sum_{s=0}^{3}(\mathbf{B}_i(s) * (y_i-\mu_s))
\end{aligned}$$

Then,

$$\frac{\partial \ell}{\partial \mu_s} = \sum_{i=1}^{n}\frac{\mathbf{B}_i(s)}{\mathbf{A}_i}$$

where $s = 0, 1, 2, 3$.

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{\mathbf{C}_i}{\mathbf{A}_i}$$

3.5.2 The Second Derivatives of the Loglikelihood Function

We define the following notations:

$$\begin{aligned}
\mathbf{B2}_i(s,s) &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial \mu_s^2} = \frac{\partial \mathbf{B}_i(s)}{\partial \mu_s} \\
&= \mathbf{B}_i(s)\frac{y_i-\mu_s}{\sigma^2} - \frac{\mathbf{B}_i(s)}{y_i-\mu_s} \\
&= \mathbf{B}_i(s)(\frac{y_i-\mu_s}{\sigma^2} - \frac{1}{y_i-\mu_s})
\end{aligned}$$

where $s = 0, 1, 2, 3$.

$$\begin{aligned}\mathbf{D}_i(s) &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial \mu_s \partial \sigma^2} = \frac{\partial \mathbf{C}_i}{\partial \mu_s} \\ &= \frac{1}{2\sigma^2}(\mathbf{B2}_i(s,s)(y_i - \mu_s) - \mathbf{B}_i(s))\end{aligned}$$

where, $s = 0, 1, 2, 3$.

and

$$\begin{aligned}\mathbf{E}_i &\doteq \frac{\partial^2 \mathbf{A}_i}{\partial (\sigma^2)^2} = \frac{\partial \mathbf{C}_i}{\partial \sigma^2} \\ &= \sum_{s=0}^{3}(\frac{1}{2\sigma^2}\mathbf{D}_i(s) * (y_i - \mu_s) - \frac{1}{2(\sigma^2)^2}\mathbf{B}_i(s) * (y_i - \mu_s))\end{aligned}$$

Then, the second derivatives are

$$\frac{\partial^2 \ell}{\partial \mu_s^2} = \sum_{i=1}^{n}(\frac{\mathbf{B2}_i(s)}{\mathbf{A}_i} - \frac{\mathbf{B}_i(s)^2}{\hat{\mathbf{A}_i^2}}), (s = 0, 1, 2, 3)$$

$$\frac{\partial^2 \ell}{\partial \mu_s \partial \mu_t} = \sum_{i=1}^{n}(-\frac{\mathbf{B}_i(s)\mathbf{B}_i(t)}{\hat{\mathbf{A}_i^2}}), (s, t = 0, 1, 2, 3; s \neq t)$$

$$\frac{\partial^2 \ell}{\partial \mu_s \partial \sigma^2} = \sum_{i=1}^{n}(\frac{\mathbf{D}_i(s)}{\mathbf{A}_i} - \frac{\mathbf{C}_i\mathbf{B}_i(s)}{\hat{\mathbf{A}_i^2}}), (s = 0, 1, 2, 3)$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} + \sum_{i=1}^{n}(\frac{\mathbf{E}_i}{\mathbf{A}_i} - \frac{\mathbf{C}_i^2}{\hat{\mathbf{A}_i^2}})$$

BIBLIOGRAPHY

Andersson L., Haley C.S., Ellegren H., Knott S.A., Johansson M., Andersson K., Andersson-Eklund L., Edfors-Lilja I., Fredholm M., Hansson I. and et al. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263**, 1771–1774.

Azevedo-Filho A. and Shachter R.D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. *In Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference* pp. 28–36.

Beavis W.D., Grant D., Albertsen M. and Fincher R. (1991). Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. *Theoretical and Applied Genetics* **83**, 141–145.

Berrettini W.H., Ferraro T.N., Alexander R.C., Buchberg A.M. and Vogel W.H. (1994). Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. *Nature Genetics* **7**, 54–58.

Berry C.C. (1998). Computationally efficient bayesian qtl mapping in experimental crosses. *ASA Proceedings of the Biometrics Section* pp. 164–169.

Brem R.B., Yvert G., Clinton R. and Kruglvak L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755.

Bystrykh L., Weersing E., Dontje B., Sutton S., Pletcher M.T., Wiltshire T., Su A.I., Vellenga E., Wang J., Manly K.F., Lu L., Chesler E.J., Alberts R., Jansen R.C., Williams R.W., Cooke M.P. and Haan G.d. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* **37**, 225–232.

Carlin B.P. and Chib S. (1995). Bayesian model choice via markov chain monte carlo. *J. Am. Stat.* **88**, 881–889.

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

Cowen N.M. (1989). Multiple linear regression analysis of rflp data sets used in mapping qtls. *Development of application of molecular markers to problems in plant genetics, Cold Spring Harbor, New York* pp. 113–116.

Crawford S.L. (1994). An application of the laplace method to finite mixture distributions. *Journal of The American Statistical Association* **89**, 259–267.

Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B* **39(1)**, 1–38.

deVicente M.C. and Tanksley S. (1993). Qtl analysis of transgressive segregation in an interspecific tomato cross. *Genetics* **134**, 585–596.

Doerge R.W. and Churchill G.A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Doerge R.W., Zeng Z.B. and Weir B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **13**, 195–219.

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

Gaffney P.J. (2001). An efficient reversible jump markov chain monte carlo approach to detect multiple loci and their effects in inbred crosses. *Ph.D. Dissertation, Department of Statistics, University of Wisconsin, Madison, WI.* .

Gelfond J.A.L., Ibrahim J.G. and Zou F. (2007). Proximity model for expression quantitative trait loci (eqtl) detection. *Biometrics* .

Green P.J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82**, 711–732.

Haldane J.B.S. and Waddington C.H. (1931). Inbreeding and the linkage. *Genetics* **16**, 357–374.

Haley C.S. and Knott S.A. (1992). A simple regresssion method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* .

Hastings W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109.

Huang H., Eversley C.D., Threadgill D.W. and Zou F. (2007). Bayesian multiple quantitative trait loci mapping for complex traits using markers of the entire genome. *Genetics* .

Hubner N., Wallace C.A., 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., Mu"ller A., Cook S.A., Kurtz T.W., Whittaker J., Pravenec M. and Aitman T.J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**, 243–253.

Ishimori N., Li R., Kelmenson P.M., Korstanje R., Walsh K.A., Churchill G.A., Forsman-Semb K. and Paigen B. (2004). Quantitative trait loci analysis for plasma hdl-cholesterol concentrations and atherosclerosis susceptibility between inbred mouse strains c57bl/6j and 129s1/svimj. *Arterioscler Thromb Vasc Biol* .

Jansen R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* .

Jansen R.C. and Stam P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* pp. 1447–1455.

Kao C.H. and Zeng Z.B. (1997). General formulas for obtaining the mles and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the em algorithm. *Biometrics* **53**, 653–665.

Kao C.H., Zeng Z.B. and Teasdale R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* pp. 1203–1216.

Kass R.E. and Raftery A.E. (1995). Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.

Kendziorski C., Chen M., Yuan M., Lan H. and Attie A.D. (2006). Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics* **62**, 19–27.

Kirst M., Myburg A.A., De Leo'n P.G., Kirst M.E., Scott J. and Sederoff R. (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cdna microarray data in an interspecific backcross of eucalyptus. *Plant Physiol.* **135**, 2368–2378.

Knapp S.J. (1991). Using molecular markers to map multiple quantitative trait loci: models for backcroos, recombinant inbred, and doubled haploid progeny. *Theoretical and Applied Genetics* **81**, 333–338.

Kong A. and Wright F.A. (1994). Asymptotic theory for gene mapping. *Proceedings of the National Academy of Sciences* **91**, 9705–9709.

Lander E.S. and Bostein D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics* pp. 185–199.

Lee M.T., Kuo F.C., Whitemore G.A. and Sklar J. (2000). Importance of replication in microarray expression studies: statistical methods and evidence from repetitive cdna hybridizations. *Proceedings of the National Academy of Sciences USA* **97**, 9834–9839.

Lipshutz R.J., Fodor S., Gingeras T.R. and Lockhart D.J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**, 20–24.

Lockhart D.J., Dong H., Byrne M.C., Follettie M.T., Gallo M.V., Chee M.S., Mittmann M., Wang C., Kobayashi M., Horton H. and Brown E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1676–1680.

Martinez O. and Curnow R.N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Meng X.L. and Rubin D.B. (1993). Maximum likelihood estimation via the ecm algorithn: a general framework. *Biometrika* **80**, 267–278.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Miller A.J. (1990). Subset selection in regression .

Monks S.A., Leonardson A., Zhu H., Cundiff P., Pietrusiak P., Edwards S., Phillips J.W., Sachs A. and Schadt E.E. (2004). Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105.

Morley M., Molony C.M., Weber T.M., Devlin J.L., Ewens K.G., Spielman R.S. and Cheung V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

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

Ott J. (1991). Analysis of human genetic linkage .

Satagopan J.M. and Yandell B.S. (1996). Estimating the number of quantitative trait loci via bayesian model determination. *In Proceedings of the Section on Biometrics. Joint Statistical Meetings, Chicago* .

Satagopan J.M., Yandell B.S., Newton M.A. and Osborn T.G. (1996). A bayesian approach to detect quantitative trait loci using markov chain monte carlo. *Genetics* pp. 805–816.

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

Sillanpaa M.J. and Arjas E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* pp. 1373–1388.

Simpson S.P. (1989). Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **77**, 815–819.

Soller M., Brody T. and Genizi A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Stam P. (1991). Some aspects of qtl analysis. *in Proceedings of the English Meeting of the Eucarpia Section Biometrics in Plant Breeding. Brno* .

Stephens D.A. and Fisch R.D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump markov chain monte carlo. *Biometrics* pp. 1334–1347.

Stuber C.W., Lincoln S.E., Wolff D.W., Helentjaris T. and Lander E.S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Theoretical and Applied Genetics* **132**, 823–839.

Sugiyama F., Churchill G.A., Higgins D.C., Johns C., Makaritsis K.P., Gavras H. and Paigen B. (2001). Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics* **71**, 70–77.

Thisted R.A. (Mar. 1998). Elements of statistical computing: Numerical computation. *CRC Press* .

Thoday J.M. (1961). Location of polygens. *Nature* **191**, 368–370.

Thomas D.C., Richardson S., Gauderman J. and Pitkaniemi J. (1997). A bayesian approach to multipoint mapping in nuclear families. *Genet. Epidemiol.* **14**, 903–908.

Venables W.N. and Ripley B.D. (1994). Modern applied statistics with s-plus. *Springer.* pp. 105–110.

Weller J.I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

Weller J.I. (1987). Mapping and analysis of quantitative trait loci in lycopersicon (tomato) with the aid of genetic markers using approximate maximum likelihood methods. *Heredity* **59**, 413–421.

Wright F.W. and Kong A. (1997). Linkage mapping in experimental crosses: the robustness of single-gene models. *Genetics* **146**, 417–425.

Xu S. (1996). Mapping quantitaive trait loci for complex binary diseases using line crosses. *Genetics* **143**, 1471–1424.

Yi N. (2004). A unified markov chain monte carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**, 967–975.

Yi N. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**, 1333–1344.

Yi N., Allison D.B. and Xu S. (2003). Bayesian model choice and search strategies for mapping multiple epistatic quantitative trait loci. *Genetics* **165**, 867–883.

Yi N. and S. X. (2000). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* pp. 1391–1403.

Yi N., Shriner D., Banerjee S., Mehta T., Pomp D. and Yandell B.S. (2007a). An efficient bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**, 1865–1877.

Yi N. and Xu S. (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* pp. 1759–1771.

Yi N. and Xu S. (2002). Mapping quantitative trait loci with epistatic effects. *Genet. Res.* **79**, 185–198.

Yi N., Yandell B.S., Churchill G.A., Allison D.B., Eisen E.J. and Pomp D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**, 1333–1344.

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

Zeng Z.B. (1993). Theoretical basis of crosses between inbred strains of gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* .

Zeng Z.B. (1994). Precision mapping of quantitative trait loci. *Genetics* .

Zeng Z.B. (2000). Statistical methods for mapping quantitative trait loci. *In preparation* p. 9.