BAYESIAN FUNCTIONAL MAPPING
OF COMPLEX DYNAMIC TRAITS

By
TIAN LIU

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
UNIVERSITY OF FLORIDA
2007
To my mum, my husband Shenghua, and my little sweetheart DuDu.
ACKNOWLEDGMENTS

There are lots of people I would like to thank for a huge variety of reasons.

Firstly, I am grateful to my thesis advisor, Dr. Rongling Wu, for his guidance and support. I could not have imagined having a better advisor and mentor for my PhD, and this dissertation would not have been possible without his knowledge, perceptiveness and expert guidance. Not only was he readily available for me, as he so generously is for all of his students, but he always read and responded to the drafts of each chapter of my work more quickly than I could have hoped. I was in the classes of Dr. Malay Ghosh and Dr. Ramon Littell and I really appreciate their kindness and considerable mentoring. Many thanks also go to committee members Dr. Xueli Liu and Dr. Hartmut Derendorf. In every sense, none of this work would have been possible without their helps.

I also thank all the rest of the academic and support staff of the Department of Statistics at the University of Florida. Much gratitude is given to my officemates, and friends, Hongying and Song for putting up with me for almost three years.

A penultimate thank-you goes to my mother and my mother-in-law. For always being there when I needed them most, and helping me to take good care of my new-born baby, they deserve far more credit than I can ever give them.

My final, and most heartfelt, acknowledgment must go to my husband Shenghua. His support, encouragement, and companionship has turned my journey through graduate school into a pleasure. For all that, and for being everything I am not, he has my everlasting love.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>7</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>11</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td><strong>1</strong> INTRODUCTION</td>
<td>13</td>
</tr>
<tr>
<td>1.1 Genetics of Quantitative Traits</td>
<td>13</td>
</tr>
<tr>
<td>1.2 Nature of Quantitative Variation</td>
<td>15</td>
</tr>
<tr>
<td>1.3 Genetic Mapping: from Static to Dynamic</td>
<td>15</td>
</tr>
<tr>
<td>1.4 Mapping Approaches: from Frequentist to Bayesian</td>
<td>17</td>
</tr>
<tr>
<td>1.5 Dissertation Goals</td>
<td>20</td>
</tr>
<tr>
<td><strong>2</strong> A GENERAL BAYESIAN-FUNCTIONAL MAPPING FRAMEWORK</td>
<td>21</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>21</td>
</tr>
<tr>
<td>2.2 Functional Mapping</td>
<td>24</td>
</tr>
<tr>
<td>2.2.1 Linear Model</td>
<td>24</td>
</tr>
<tr>
<td>2.2.2 Modeling the Mean-Covariance Structures</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Likelihood</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Parameter Estimation</td>
<td>28</td>
</tr>
<tr>
<td>2.5 Implementation</td>
<td>29</td>
</tr>
<tr>
<td>2.6 Estimation Issues</td>
<td>32</td>
</tr>
<tr>
<td>2.7 Modeling the Covariance Matrix by a Reference Prior</td>
<td>33</td>
</tr>
<tr>
<td>2.8 Modeling the Structure of the Covariance Matrix</td>
<td>35</td>
</tr>
<tr>
<td>2.8.1 Modeling the Structure of the Within-Subject Covariance Matrix</td>
<td>35</td>
</tr>
<tr>
<td>by the First-Order Autoregressive Model</td>
<td>35</td>
</tr>
<tr>
<td>2.8.2 Modeling the Structure of the Within-Subject Covariance Matrix</td>
<td>38</td>
</tr>
<tr>
<td>by the Structured Antedependence Model</td>
<td>38</td>
</tr>
<tr>
<td>2.9 A Worked Example</td>
<td>40</td>
</tr>
<tr>
<td>2.9.1 Animal Material</td>
<td>40</td>
</tr>
<tr>
<td>2.9.2 Results</td>
<td>40</td>
</tr>
<tr>
<td>2.10 Monte Carlo Simulation</td>
<td>46</td>
</tr>
<tr>
<td>2.11 Discussion</td>
<td>53</td>
</tr>
<tr>
<td><strong>3</strong> A BAYESIAN MODEL FOR MAPPING EPISTATIC QTLS THAT REGULATE DYNAMIC PROCESSES</td>
<td>59</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>59</td>
</tr>
<tr>
<td>3.2 Model</td>
<td>60</td>
</tr>
</tbody>
</table>
3.3 A Worked Example ........................................... 63
  3.3.1 Material .................................................. 63
  3.3.2 Results ................................................... 63
3.4 Monte Carlo Simulation ....................................... 69
3.5 Discussion .................................................... 82

4 PROSPECTS ....................................................... 83
  4.1 Introduction .................................................. 83
  4.2 Parametric, Nonparametric and Semiparametric Modeling .......... 83
  4.3 Toward a Comprehensive Biology ............................... 85
    4.3.1 Joint Models of Longitudinal Trajectories and Time-to-Events ... 85
    4.3.2 Multivariate Longitudinal Trajectories ....................... 86
  4.4 Statistical Considerations .................................... 86
    4.4.1 Model Selection ......................................... 86
    4.4.2 Sensitivity Studies and Other Estimation Issues ............. 87
    4.4.3 Testing QTL Effects .................................... 87
  4.5 Conclusions .................................................. 88

REFERENCES ...................................................... 90

BIOGRAPHICAL SKETCH ........................................... 98
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>The sample variance-covariance matrix for the mice-body-mass data.</td>
<td>56</td>
</tr>
<tr>
<td>2-2</td>
<td>The sample variance-covariance matrix for the transformed mice-body-mass data.</td>
<td>56</td>
</tr>
<tr>
<td>2-3</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>56</td>
</tr>
<tr>
<td>2-4</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>57</td>
</tr>
<tr>
<td>2-5</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 7. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>57</td>
</tr>
<tr>
<td>2-6</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>57</td>
</tr>
<tr>
<td>2-7</td>
<td>Results from a simulation study by assuming none covariance structure. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>57</td>
</tr>
<tr>
<td>2-8</td>
<td>Results from a simulation study by assuming the covariance structure to be SAD(1). Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>58</td>
</tr>
<tr>
<td>2-9</td>
<td>Results from a simulation study by assuming the covariance structure to be AR(1). Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>58</td>
</tr>
<tr>
<td>2-10</td>
<td>Results from a simulation study by performing traditional maximum-likelihood typed method. Numbers in the box bracts are the given values of the parameters.</td>
<td>58</td>
</tr>
<tr>
<td>3-1</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6 and 7. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>66</td>
</tr>
<tr>
<td>3-2</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6 and 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>70</td>
</tr>
<tr>
<td>3-3</td>
<td>Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 7 and 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.</td>
<td>71</td>
</tr>
</tbody>
</table>
3-4 Estimated Bayes factors on a logarithmic scale between different models . . . . 72

3-5 Results from a simulation study, where \( s = 2 \) and \( H^2 = 0.1 \). Numbers in the parentheses are the 95% equal-tail confidence intervals. Numbers in the box bracts are the given values of the parameters. . . . . . . . . . . . . . . . . . . 75

3-6 Results from a simulation study, where \( s = 2 \) and \( H^2 = 0.4 \). Numbers in the parentheses are the 95% equal-tail confidence intervals. Numbers in the box bracts are the given values of the parameters. . . . . . . . . . . . . . . . . . . 81
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Autocorrelation function of the locus ($\lambda$) at different lags. In the first plot, the MCMC samples were subsampled at every 30th cycles. In the second plot, the MCMC samples were subsampled at every 60th cycles.</td>
<td>42</td>
</tr>
<tr>
<td>2-2</td>
<td>Trace plots for the estimated parameters.</td>
<td>43</td>
</tr>
<tr>
<td>2-3</td>
<td>A profile of Estimated marginal posterior distribution of the QTL location by assuming that exactly one QTL is located on one of the chromosome respectively.</td>
<td>44</td>
</tr>
<tr>
<td>2-4</td>
<td>Fitted growth curves for the three QTL genotypes assuming a single QTL is located on mouse chromosome 6.</td>
<td>45</td>
</tr>
<tr>
<td>2-5</td>
<td>Fitted growth curves for the three QTL genotypes assuming the QTL is located on mouse chromosome 7.</td>
<td>46</td>
</tr>
<tr>
<td>2-6</td>
<td>Fitted growth curves for the three QTL genotypes assuming the QTL is located on mouse chromosome 10.</td>
<td>47</td>
</tr>
<tr>
<td>2-7</td>
<td>Dynamic changes of the additive and dominant effect due to the QTL located on mouse chromosome 6,7, and 10 respectively.</td>
<td>48</td>
</tr>
<tr>
<td>2-8</td>
<td>Estimated marginal posterior distribution of the QTL location on mouse chromosome 6,7, and 10 respectively.</td>
<td>49</td>
</tr>
<tr>
<td>2-9</td>
<td>Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming no covariance structures.</td>
<td>50</td>
</tr>
<tr>
<td>2-10</td>
<td>Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming the covariance structure is SAD(1).</td>
<td>51</td>
</tr>
<tr>
<td>2-11</td>
<td>Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming the covariance structure is AR(1).</td>
<td>52</td>
</tr>
<tr>
<td>2-12</td>
<td>The profile of the LR between the full and reduced (no QTL) model estimated from the SAD(1) model for body mass growth trajectories in a simulation study. As to the settings of this simulation, we assume that a single QTL control body mass growth trajectories of mice.</td>
<td>53</td>
</tr>
<tr>
<td>3-1</td>
<td>Estimated marginal posterior distribution of the QTL location on mouse chromosome 6,7, and 10 respectively.</td>
<td>65</td>
</tr>
<tr>
<td>3-2</td>
<td>Estimated marginal posteriors of the two loci based on a two-QTL-model.</td>
<td>68</td>
</tr>
<tr>
<td>3-3</td>
<td>Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 6 and 7.</td>
<td>69</td>
</tr>
</tbody>
</table>
3-4 Dynamic changes of the additive, dominant and interaction effects due to the two QTL based on a tow-QTL-model. .......................... 70

3-5 Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 6 and 10. .......................... 71

3-6 Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 7 and 10. .......................... 72

3-7 Estimated marginal posterior distributions of the two loci on a same chromosome in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.1. 77

3-8 The profile of the LR between the full and reduced (no QTL) model estimated from the SAD(1) model for body mass growth trajectories in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.1. 78

3-9 Estimated marginal posterior distributions of the two loci on a same chromosome in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.4. 79

3-10 Fitted and given growth curves for the 9 QTL genotypes in a simulation study, assuming the heritability is 0.4. .......................... 80
Many quantitative traits of fundamental importance to agricultural, evolutionary, and biomedical genetic research can better be described as dynamic processes. Understanding the genetic control of such dynamic or longitudinal traits (such as growth curves, HIV dynamics and drug response) has been a long-standing challenge because of their intrinsic developmental complexity. More recently, a general statistical framework, called functional mapping, has been proposed to map quantitative trait loci (QTLs) that regulate the developmental pattern and process of dynamic traits. Functional mapping has proven to be biologically relevant because it is incorporated by fundamental biological principles to test the genetic and developmental mechanisms for trait changes based on tractable mathematical functions. Original functional mapping was derived within the maximum likelihood (ML) context and implemented with the EM algorithms. Although ML-based functional mapping has many favorable statistical properties for parameter estimation, it has quickly become limited in capacity when a high-dimensional longitudinal problem, as commonly seen in systems biology, is encountered.

In my research, I derive a general functional mapping framework for QTL mapping of dynamic traits within the Bayesian paradigm. The Markov Chain Monte Carlo (MCMC) techniques were implemented for functional mapping to estimate biologically and statistically sensible parameters that model the structures of time-dependent genetic effects and covariances. The Bayesian approach is useful to handle difficulties
in constructing confidence intervals as well as the identifiability problem, enhancing the statistical inference of functional mapping. By comparing the Bayes factors from separate models, the actual number of QTLs that are involved in the dynamic variation of a trait can be estimated. The model framework was extended to estimate the effects of epistatic interactions between different QTLs on dynamic traits in various developmental stages. I undertaken extensive simulation studies to investigate the statistical behavior of the new statistical model and used a real example for the F$_2$ mice to validate model utilization. Bayesian-based functional mapping via MCMC algorithms estimates parameters that determine the shape and function of a particular biological process, thus providing a flexible platform to test biologically meaningful hypotheses regarding the complex relationships between gene actions or interactions and developmental processes.
Genetics has been thought to play a dominant role in explaining fundamental biological issues in life sciences, given that almost every biological phenomenon involves a genetic component. As a result of this, there is a pressing need for studying the genetic architecture of biological traits, especially those that vary in continuous patterns. The past two decades have witnessed unprecedented growth in our ability to dissect these so-called quantitative traits into individual loci at the molecular level in agricultural genetics, evolutionary biology [65] [66] [93] [103] and human genetics aimed to detect genes for complex human diseases [67]. Quantitative traits are thought to be controlled by multiple genes, each with a small effect and segregating according to Mendel’s laws, and can also be affected by the environment to varying degrees [53]. According to this argument established by R. A. Fisher [24], the observed phenotype of a quantitative trait \( y \) can be expressed as a linear combination of genetic \( g \), environmental \( e \) and genotype × environment interaction effects, i.e.,

\[
y = \mu + g + e + (g \times e) + \epsilon,
\]

where \( \mu \) is the population mean and \( \epsilon \) is the residual error. Depending on different purposes of plant breeding, the environmental effect can be due to different climates or locations [69], which are usually called “macroenvironments” in light of their evident varying patterns [102]. The macroenvironment effect can be discrete, like location, or continuous, such as temperature, moisture and nutrient, in nature [91]. The residual error is due to stochastic fluctuations, i.e., “microenvironments” [102]. Under the prerequisite that the families or genotypes studied have multiple replicates in space, statistical approaches based on regression models and analysis of variance have been available to estimate the variances due to the genetic, environment and residual effects, as
well as genotype environment interaction effects. Heritability, defined as the proportion of the genetic variance to the total phenotypic variance, is then estimated to measure the contribution of genetic factors to quantitative variation.

With the development of modern molecular markers, classical quantitative genetics has been developed to a point at which individual genetic loci underlying a quantitative trait, called quantitative trait loci (QTLs), can be mapped and identified [53]. By dissecting a quantitative trait into a total of \( m \) possible QTLs each segregating in a Mendelian ratio, the phenotypic value of the trait can be expressed as

\[
y = \mu + \sum_{r=1}^{m} g_r + \sum_{r \neq s} (g_r \times g_s) + e + \sum_{r=1}^{m} (g_r \times e) + \sum_{r \neq s} (g_r \times g_s \times e) + \epsilon,
\]

where \( \sum_{r=1}^{m} \) and \( \sum_{r \neq s} \) denote the summations associated with the main and epistatic genetic effects, respectively, among the QTLs estimated from a linkage map constructed by molecular markers. Equation 1–2 presents a general statistical model for QTL mapping [35] [42] [45] [46] [49]. The detection of the underlying QTL for a quantitative trait is based on a segregating population of progeny derived from crossing genotypes containing different alleles at phenotypically important loci. The crossed parents should be adequately divergent to identify discrete molecular markers that sample the genome at sufficiently dense intervals. Statistical principles and methods for mapping QTLs with the linkage map constructed from genotyped molecular markers have been well established [46] [50] [54] [69] [70] [107]. In practice, QTL mapping approaches have been instrumental for the characterization and discovery of thousands of QTLs responsible for a variety of traits in plants, animals and humans. In a recent study, Li et al. [52] were able to characterize the molecular basis of the reduction of grain shattering – a fundamental selection process for rice domestication – at the detected QTL. Many other examples for the success of QTL mapping include the positional cloning of QTL responsible for fruit size and shape in tomato [26] and for branch, florescence and grain architecture in maize [19] [28] [98].
1.2 Nature of Quantitative Variation

Most quantitative traits are determined by a web of many interacting loci and by an array of environmental factors [22]. The traditional polygenic theory of quantitative traits [57] envisaged a fairly large number of loci, each with relatively small and equal effects, acting in a largely additive way. Over the years it has indeed been observed that a quantitative trait may display complicated genetic architecture [2] [56], described below:

1. It may be controlled by a fairly large number of loci; for example, of the order of 50, according to the work of Shrimpton and Robertson [79] [80];
2. Genes act in ways which may be additive, dominant, epistatic and interactive with environmental factors;
3. The magnitude of the effect produced by each locus can vary considerably;
4. The same genes may affect different phenotypic traits through pleiotropic effects;
5. The genes affecting the trait may be distributed over the genome at random or in a certain pattern.

With the use of genetic mapping to analyze quantitative traits, increasing evidence has been observed for the third point, which suggests that typically a small number of loci account for a very large fraction of the variation in the trait. For this reason, the traditional polygenic model may be replaced by a new oligogenic model in which a small number of major genes each with a large effect, combined with many minor genes each with a small effect, determine the genetic variation of a quantitative trait (see [55] for an excellent review). According to the oligogenic model, the distribution of genetic effects may be approximated by a geometric series [50]. When incorporated into a QTL mapping model, such an approximation can significantly increase the power of QTL mapping and the precision of parameter estimation.

1.3 Genetic Mapping: from Static to Dynamic

Every living entity with an incredible range of body sizes from microbes (10-13 g) to blue whales (108,000 Kg) typically exhibits a growth and developmental pattern, i.e., its size (in terms of length, area or mass) and shape change with age or other independent variables through altering the ratio of the anabolic or metabolic rate and the rate of
catabolism. For this reason, a quantitative trait of living entities should be regarded as a live process, in order to best characterize developmental features of the trait and fully explore its relationships with other parts of the living entities.

Currently, most genetic approaches describe a biological trait by the value measured at a single time point during its ontogeny or a single state in an environmental space. This is largely insufficient and, rather, a precise description of the trait should include a series of measurements taken at multiple discrete time points or states. Empirical analyses of time-dependent growth data suggest that growth trajectories (changes of the trait value over time) follow particular exponential laws that can be mathematically elucidated by the so-called growth functions [4]. Based on fundamental principles behind biological or biochemical networks, [99] have mathematically proven the universality of these growth equations. Apart from the growth law, many mathematical equations have been well constructed to describe various life processes of paramount importance in agricultural, biomedical and health sciences [38] [68].

The genetic mapping of a quantitative trait expressed as a trajectory or curve presents one of the most challenging issues in genetic research because of the infinite-dimensional [48] or function-valued nature of the curve [71]. However, some of key difficulties in mapping have been overcome by R. Wu and colleagues ([54] [104] [105] [106], and reviewed in [107]). They have proposed a general statistical framework, i.e., functional mapping, to genome-wide map specific QTL that determine the developmental pattern of a complex trait.

The basic rationale of functional mapping lies in the connection between gene action or environmental effects and development by parametric or nonparametric models. Functional mapping maps dynamic QTL that are responsible for a biological process that is measured at a finite number of time points through established mathematical models for defining the developmental process of a biological phenotype. With mathematical functions incorporated into the QTL mapping framework, functional mapping estimates
parameters that determine shapes and functions of a particular biological network, instead of directly estimating the gene effects at all possible time points. Because of such connections among these points through mathematical functions, functional mapping strikingly reduces the number of parameters to be estimated and, hence, displays increased statistical power.

From a statistical perspective, functional mapping is a problem of jointly modelling mean-covariance structures in longitudinal studies, an area that has recently received a considerable interest in the statistical literature \[64\] \[72\] \[73\] \[101\]. However, different from general longitudinal modelling, functional mapping integrates the parameter estimation and test process within a biologically meaningful mixture-based likelihood framework. Functional mapping is thus advantageous in terms of biological relevance because biological principles are embedded into the estimation process of QTL parameters. The results derived from functional mapping will be closer to biological reality.

1.4 Mapping Approaches: from Frequentist to Bayesian

Functional mapping was constructed within the maximum likelihood context, incorporated by a finite mixture model. For a mixture model, each observation is assumed to have arisen from one of the known or unknown number of components, each component being modelled by a density from the parametric family. Assuming that there are \(J\) QTL genotypes contributing to growth trajectories measured at \(T\) time points (denoted by \(y\)), this mixture model is expressed as

\[
y \sim p(y|\varpi, \Omega_u, \Omega_v) = \varpi_1 f_1(y; \Omega_{u1}, \Omega_v) + \cdots + \varpi_J f_J(y; \Omega_{uJ}, \Omega_v),
\]

(1–3)

where \(\varpi = (\varpi_1, \cdots, \varpi_J)\) are the mixture proportions (i.e., QTL genotype frequencies) which are constrained to be non-negative and sum to unity, \(\Omega = (\Omega_{u1}, \cdots, \Omega_{uJ})\) is a vector that contains the parameters specific to component (or QTL genotype) \(j\) and \(\Omega_v\) includes the parameters common to all components. Each density is modelled by a multivariate normal distribution.
The maximum likelihood method for parameter estimation in a mixture model obtains point estimates \((\hat{\varpi}, \hat{\Omega}, \hat{\Omega}_v)\) of the parameters \(\Theta = (\varpi, \Omega, \Omega_v)\) by maximizing the likelihood of equation 1–3. The maximum likelihood estimates (MLEs) of the unknown parameters \(\Omega\) under the mixture QTL model can be computed by implementing an EM algorithm [16] [58]. However, there can sometimes be problems when ML is used for QTL mixture models. First, for some choices of parametric families, \(f\), the likelihood is unbounded. Second, in complex situations the likelihood function can have many local maxima, each of which may give different (and possibly reasonable) plug-in estimates for quantities of interest. In these cases it could be difficult in choosing one of these point estimates of the parameters above the others. Third, to perform significance tests and obtain confidence interval estimates of the estimators, substantial computation on repeated sampling through permutation tests [12] or bootstrapping [92] is required. Furthermore, these approaches do not properly account for uncertainties in the other parameters, making it unreliable to claim coverage probabilities of the confidence intervals.

Fourth and most importantly, when a QTL mapping strategy is incorporated by autocorrelated longitudinal data, we will encounter a much higher dimensional space for the unknown parameters than traditional QTL modelers. Although the mere existence of a high-dimensional parameter space is not necessarily detrimental, extra care must be taken in searching for the ML estimator. An extra complication (not only of ML) is that the uncertainty about the actual number of QTL for a longitudinal quantitative trait results in extra difficulty in model fitting and selection.

The Bayesian method can avoid many of the problems described above for ML. In ML, the unknown parameters are treated as unknown variables (unobservables) and the likelihood function is maximized in these variables. In the Bayesian paradigm, each unobservable parameter is given a prior distribution, and we then infer the posterior distribution of each unobservable conditional on the data (the observables). The summary statistics of the posterior distribution, e.g., the mean, the mode or the median, can be
regarded as Bayesian estimates of unobservables [8]. The interval estimate can be obtained simply by examining the posterior distribution. Let us denote the observables by a vector \( y \) (the data vector) and unobservables by a parameter vector \( \Theta = (\lambda, \Omega, \Omega_\nu) \). The posterior distribution is

\[
p(\Theta|y) = \frac{p(y, \Theta)}{p(y)} = \frac{p(y|\Theta)p(\Theta)}{p(y)} \propto p(y|\Theta)p(\Theta)
\]

where \( p(*) \) is a generic expression for a probability density, \( p(y|\Theta) \) is the likelihood and \( p(\Theta) \) is the prior probability distribution of the unobservables. Because the dominator is just the probability density of \( y \), not a function of the parameters, it can be ignored.

We partition the vector \( \Theta \) into \( \Theta = [\Theta_\ell \{\Theta\}_-\ell] \) where \( \Theta_\ell \) is a single element of the unobservables and \( \Theta_-\ell \) is the rest of the unobservables that exclude \( \Theta_\ell \). The marginal posterior distribution of \( \Theta_\ell \) is expressed by

\[
p(\Theta_\ell|y) = \int p(\Theta_\ell, \Theta_-\ell|y)d\Theta_-\ell \propto \int p(y|\Theta_\ell, \Theta_-\ell)p(\Theta_\ell, \Theta_-\ell)d\Theta_-\ell
\]

The mean of this marginal posterior distribution is a candidate Bayesian estimator of \( \Theta_\ell \). This marginal distribution rarely has an explicit form, and numerical integration is often prohibited because the dimensionality of \( \Theta_-\ell \) may be high. However, a Markov chain Monte Carlo (MCMC) algorithm can be used to simulate the sample from the joint posterior distribution [77]. The potential of the Bayesian approach implemented with the Gibbs sampler or Metropolis-Hastings algorithm for genome mapping has been explored for several relatively simple genetic designs [78] [81]. In particular, because we are now able to examine the entire posterior distribution of each parameter, we will be better able to deal with problems such as multi-modality of the likelihood function. Baysian inference for gene mapping was first introduced by [39], [40], and [86] [88]. Models for implementing multiple QTL were developed by Satagopan et. al. [78], Heath [37], Uimari and Hoeschele [89], Stephens and Fisch [85], Sillanpaa and Arjas [81] [82]. The determination of the
actual number of QTL is made by comparing the Bayes factors from separate models or by applying a reversible jump algorithm.

1.5 Dissertation Goals

To my knowledge, there is no reference thus far that deals with the Bayesian genetic mapping of QTL controlling dynamic traits. In this dissertation, I will develop a general Bayesian framework for functional mapping of complex dynamic traits based on parametric modelling of the mean-covariance structures. This framework is constructed by a mixture model in which multiple mixture components corresponding to the genotypes of the underlying QTL are involved. I will implement the MCMC algorithm to estimate the posterior distribution of each parameter contained within the mixture model. More specifically, I will focus on the following aspects of Bayesian functional mapping:

1. Incorporate biologically meaningful parameters that define the growth curve of a dynamic trait into functional mapping, and then derive a Bayesian procedure for their estimation;

2. Model the structure of covariance matrix among time-dependent observations with commonly used statistical approaches, such as the stationary first-order autoregressive, AR(1), or nonstationary first-order structured antedependence, SAD(1), and implement a Bayesian approach to estimate the covariance-structuring parameters;

3. Derive a general approach for the genomewide enumeration of QTL that control a dynamic trait within the Bayesian paradigm;

4. Extend the Bayesian approach to estimate epistatic interactions of QTL for a dynamic trait, providing a quantitative framework for testing the role of epistasis in development.

I will perform extensive simulation studies to investigate the statistical properties of my Bayesian functional mapping model in terms of its convergence rate, estimation precision and power for QTL detection. A real example for the F₂ mouse progeny will be used to demonstrate the utilization of the model and validate its usefulness in a practical genomic project of dynamic QTL.
CHAPTER 2
A GENERAL BAYESIAN-FUNCTIONAL MAPPING FRAMEWORK

2.1 Introduction

Because of the involvement of many genes, whose expressions are often modified by and various developmental and environment factors, the outcome of many biological traits of economical, biological and biomedical importance vary in a quantitative way. The past two decades have witnessed tremendous developments of experimental and statistical approaches to dissect these quantitatively inherited traits into individual loci at the molecular level (known as quantitative trait loci or QTLs) [49] [65] [66] [67] [103]. However, because most of these approaches ignore the developmental or dynamic feature of a quantitative trait, their application to reveal the genetic and developmental basis for trait variation will be very limited.

The QTL mapping of complex dynamic traits can now be made possible with a new statistical approach, called functional mapping [54] [107]. Functional mapping incorporates fundamental biological principles behind trait growth and development into a mapping framework. It capitalizes on the mathematical aspects of a ubiquitous growth law that every biological trait experiences a growth and developmental change with time through altering the ratio of the anabolic or metabolic rate and the rate of catabolism [4] [99]. The advantages of this approach lie in its increased biological relevance by embedding biological principles into the estimation process and its flexibility to generate a number of testable hypotheses about the developmental and genetic regulation of growth in a quantitative way. From a statistical perspective, functional mapping estimates parameters that determine the shape of a genotype-specific growth curve and/or the covariance structure, instead of directly estimating individual time-dependent elements in the mean vectors and covariance matrix, which thus strikingly reduces the number of parameters to be estimated and increase the statistical power of functional mapping.
Original functional mapping was derived within the maximum likelihood (ML) context, incorporated by a finite mixture model and implemented with the EM algorithm [15], [58]. Although ML-based approaches have many favorable statistical properties for parameter estimation, they have also some significant problems when applied to a practical data set. First, the likelihood of mixture models is unbounded for some choices of parametric families. Second, there may be many local maxima for the likelihood function in some complex situations, which leads to many different different (and possibly reasonable) plug-in estimates for quantities of interest. It is difficult in choosing an optimal set of point estimates of the parameters from all possible estimates. Third, substantial computation on repeated sampling through permutation tests [14] or bootstrapping [92] is required to determine significance tests. It is also computationally expensive to obtain confidence interval estimates of the estimators. Furthermore, these approaches do not properly account for uncertainties in the other parameters, making it unreliable to claim coverage probabilities of the confidence intervals.

All the problems related to ML-based approaches above will become more serious when we deal with functional mapping of autocorrelated longitudinal data, in which a much higher dimensional space for the unknown parameters than traditional QTL modelers need to be faced. Because the parameters that model the mean and covariance structures are related in a nonlinear mathematical form, it is extremely difficult or impossible to derive the long-likelihood equations for these parameters. The last, but not only ML-specific, issue is about the actual number of QTL involved in a longitudinal quantitative trait. In ML-based QTL mapping, Kao et al [45] adopted variable selection via stepwise regression, but this has been shown to be highly computationally expensive for model fitting and selection.

Many of the problems described above for ML can be avoided by using the Bayesian method. In section 1.4, the differences in model formulation and interpretation between the ML and Bayesian approaches were described. Some key issues are highlighted here for
a better understanding of the motivation of this work. In ML, the unknown parameters are treated as unknown variables (unobservables) and the likelihood function is maximized in these variables. In the Bayesian paradigm, each unobservable parameter is given a prior distribution, and we then infer the posterior distribution of each unobservable conditional on the data (the observables). The summary statistics of the posterior distribution, e.g., the mean, the mode or the median, can be regarded as Bayesian estimates of unobservables [8]. The interval estimate can be obtained simply by examining the posterior distribution. The mean of this marginal posterior distribution is a candidate Bayesian estimator of an unknown parameter. This marginal distribution rarely has an explicit form, and numerical integration is often prohibited because the dimensionality of parameters may be high. However, a Markov chain Monte Carlo (MCMC) algorithm can be used to simulate the sample from the joint posterior distribution [77]. The potential of the Bayesian approach implemented with the Gibbs sampler or Metropolis-Hastings algorithm for genome mapping has been explored for several relatively simple genetic designs [78] [82]. Models for implementing multiple QTL were developed by Satagopan et al. [78], Heath [37], Uimari and Hoeschele [89], Stephens and Fisch [85], Sillanpaa and Arjas [81] [82]. The determination of the actual number of QTL is made by comparing the Bayes factors from separate models or by applying a reversible jump algorithm.

In this chapter, I will develop a general Bayesian framework for functional mapping of complex dynamic traits based on parametric modeling of the mean-covariance structures. This framework is constructed by a mixture model in which multiple mixture components corresponding to the genotypes of the underlying QTL are involved. I will implement the MCMC algorithm to estimate the posterior distribution of each parameter contained within the mixture model. We will perform extensive simulation studies to investigate the statistical properties of my Bayesian functional mapping model in terms of its convergence rate, estimation precision and power for QTL detection. A real example for the F₂ mouse progeny will be used to demonstrate the utilization of the model and validate
its usefulness in a practical genomic project of dynamic QTL. To clearly describe my Bayesian model, I will start with a simplest case in which only one QTL is assumed to affect a dynamic trait. The theory for such a one-QTL model can be extended to deal with any number of the underlying QTL and further draw a detailed picture of the genetic network composed of main and interactive effects on the shape and pattern of trait development.

2.2 Functional Mapping

2.2.1 Linear Model

Suppose there is a random sample of size $n$ drawn from an F$_2$ population. In this sample, multiple markers are genotyped, aimed at the identification of QTL affecting the shape of growth curves. For each individual, a particular growth phenomenon, such as body height, body weight or cell number, is measured at a series of time points ($\tau$).

Assume that there is a biallelic putative QTL with genotypes $qq(0)$, $Qq(1)$ and $QQ(2)$ affecting the shape of growth curves. At a specific time point $t$, the phenotypic value of the trait for each individual $i$ due to the QTL may be given by the linear model as follows:

$$
y_i(t) = \sum_{j=0}^{2} \xi_{ij} u_j(t) + \epsilon_i(t), \quad (t = 1, \ldots, \tau)
$$

(2–1)

where $\xi_{ij}$ is an indicator variable for individual $i$ to carry a QTL genotype and defined as 1 if a particular QTL genotype $j$ is indicated and 0 otherwise, $u_j(t)$ is the expected phenotypic value for QTL genotype $j$ at time $t$, and $\epsilon_i(t)$ is independently and identically distributed as $N(0, \sigma^2(t))$. Note that measurements within an individual are more likely to be correlated across times, with covariance $\sigma(t_1, t_2)$ between times $t_1$ and $t_2$ ($t_1, t_2 = 1, \ldots, \tau$). These variances and covariances form a $(\tau \times \tau)$ matrix $\Sigma$.

2.2.2 Modeling the Mean-Covariance Structures

Functional mapping models $u_j(t)$ by a biologically meaningful mathematical equation and the time-dependent covariance matrix composed of $\sigma^2$ and $\sigma(t_1, t_2)$ by statistically robust approaches. For example, the growth of a living entity can be defined as the
irreversible increase of size with time. A series of mathematical models have been proposed to describe the growth curves. Some early work by mathematical biologists e.g.,\cite{4} had extensive discussions on the fundamental biological aspects of mathematical modeling of growth curves. Among the various models, the sigmoidal (or logistic) growth function is regarded as being nearly universal in living systems to capture age-specific change in growth \cite{99}. Specifically, this S-shaped curve for a QTL genotype \( j \) can be mathematically expressed as:

\[
    u_j(t \mid \Omega_j) = \frac{\alpha_j}{1 + \beta_j e^{-\gamma_j t}}
\]

where \( \Omega_j = (\alpha_j, \beta_j, \gamma_j) \) is a set of parameters that determine the shape of the curves. Parameter \( \alpha \) determines the limiting value of growth as time \( t \) goes to infinity; \( \alpha/(1 + \beta) \) gives the initial value of growth, and \( \gamma \) is defined as the relative growth rate. Since the shape of the curve can be characterized by a set of parameters \( \Omega_j \), a significant difference among \( \Omega_j \) for different genotypes of a putative QTL implies that this QTL has an effect on growth curves.

A number of statistical methods have been derived to model the structure of covariances between longitudinal measurements \cite{17}. The AR(1) model has been successfully applied to model the structure of the within-subject covariance matrix for functional mapping. This model uses two simplified assumptions, i.e., variance stationarity – the residual variance (\( \sigma^2 \)) is constant over time, and covariance stationarity – the correlation between different measurements decreases proportionally (in \( \rho \)) with increased time interval.

In practice, the two simplified assumptions of the AR(1) model may not hold so that the elegant expressions of the matrix cannot be used for functional mapping. To make longitudinal data well suited to the AR(1) model, some treatments are needed. For example, in order to remove the heteroscedastic problem of the residual variance, Carroll and Rupert’s \cite{9} transform-both-sides (TBS) model is embedded into the growth-incorporated finite mixture model \cite{106}, which does not need any more parameters.
Both empirical analyses with real examples and computer simulations suggest that the TBS-based model can increase the precision of parameter estimation and computational efficiency. Furthermore, the TBS model preserves original biological means of the curve parameters although statistical analyses are based on transformed data.

The TBS-based model displays the potential to relax the assumption of variance stationarity, but the covariance stationarity issue remains unsolved. Zimmerman and Núñez-Antón [117] proposed a so-called structured antedependence (SAD) model to model the age-specific change of correlation in the analysis of longitudinal traits. The SAD model has been employed in several studies and displays many favorable properties for genetic mapping of dynamic traits [114].

2.3 Likelihood

The genetic parameters of interest are $\lambda$, $\varpi$, $\Omega$, and $\Sigma$. The parameter $\varpi$ determines the distribution of QTL genotypes in the mapping population, it is affected by the QTL position. $\lambda$ denotes the distance of the QTL from the first marker of the ordered linkage group. Assume that a linkage map has been developed for the genome. Without loss of generality, consider only one linkage group with ordered markers $1, 2, ..., m$. Also, it is assumed that the distances between markers 1 and $k$, denoted by $D = \{D_k\}_{k=1}^m$, are known. Let $M_i = \{M_{ik}\}_{k=1}^m$ denote marker genotypes of the $i$th individual.

Although, in practice, only the phenotypic values $y_i$ and the marker genotypes $M_i$ are observable, the probability distribution of the QTL genotypes for F2 populations can be expressed in terms of the location of the putative QTL, the marker genotypes and the distance between the markers. Suppose this QTL is located between markers $k$ and $k + 1$. Then, the QTL genotype distributions of the $i$th individual that reflect the mixture proportions in the mixture model 1–3 can be expressed as

$$
\varpi_{j|i} = \pi(Q_i = j | \lambda, M_{ik}, M_{ik+1}, D_k, D_{k+1}),
$$

(2–3)
The likelihood of the parameters $\lambda$, $\Omega$, and $\Sigma$ for individual $i$ can be written, after suppressing the notation for $\{M_i\}_{i=1}^{n}$ and $D$, as:

$$
L(\lambda, \Omega, \Sigma | y_i) = \sum_{j=0}^{2} \varpi_{ji} \cdot \pi(y_i | Q_i = j, \Omega, \Sigma) = \sum_{j=0}^{2} \varpi_{ji} \cdot f_j(y_i, \Omega_j, \Sigma),
$$

(2–4)

Since the data $y = \{y_i\}_{i=1}^{n}$ can be regarded as independent observations, the joint likelihood for all $n$ individuals is a product of the individual likelihoods, i.e.,

$$
L(\lambda, \Omega, \Sigma | y) = \prod_{i=1}^{n} \left[ \sum_{j=0}^{2} \varpi_{ji} \cdot \pi(y_i | Q_i = j, \Omega, \Sigma) \right].
$$

(2–5)

In the mixture model 2–5, $\Omega = \{\Omega_j\}_{j=0}^{2}$ is an unknown vector that determines the parametric family $f_j$, which presents a multivariate normal distribution with the genotype-specific mean vector expressed as

$$
\mu_j = g(t | \Omega_j) = \{\mu_j(t)\}_{t=1}^{n} = \left\{ \frac{\alpha_j}{1 + \beta_j e^{\gamma_j t}} \right\}_{t=1}^{n}
$$

(2–6)

and the covariance matrix $\Sigma$.

One aim is to make inference about $\lambda$, $\Omega$, and $\Sigma$ via this likelihood, 2–5. But when we deal with longitudinal data, it is usually not trivial to evaluate this likelihood. In addition, it is possible that not all the unknown parameters in the model 2–5 can be identified. Instead of attempting to optimize the likelihood surface, I investigated a Bayesian approach to tackle these difficulties. I will integrate the likelihood with prior knowledge to produce inference summaries for all the components in the model. Furthermore, Bayes model selection can be performed to estimate the number of QTL either by using Bayes factors or a reversible jump MCMC algorithm. The procedures for determining the number of QTL are given in the section Multiple QTL Model.
2.4 Parameter Estimation

In order to carry out the Bayesian inference, we need to specify the prior distribution. The priors could be chosen based on related studies or information from the literature. In principle, if there is reliable information for some parameters, such as Ω, priors with small dispersion may be used. For those parameters such as Σ, there is not enough prior information. Also, for the parameters of interest such as λ, noninformative priors or priors with large variance should be utilized. Given the data and the priors specified, the posterior distribution over all unknown parameters can be obtained by using the Bayes theorem. Let \( Q = \{Q_i\}_{i=1}^{n} \) denote the QTL genotypes for all \( n \) individuals. Then, the posterior density of \( \lambda, \Omega, \Sigma, \) and \( Q \) is given by

\[
\pi(\lambda, \Omega, \Sigma, Q | y) = \pi(y | Q, \Omega, \Sigma) \cdot \pi(Q | \lambda) \cdot \pi(\lambda, \Omega, \Sigma)
\]

(2.7)

where \( \pi(y | Q, \Omega, \Sigma) = \prod \pi(y_i | Q = q_i, \Omega_i, \Sigma) \) denotes the probability mass of the observation \( y \) given the QTL genotypes; \( \pi(Q | \lambda) = \prod \pi(Q_i | \lambda) \) is the probability mass of the QTL genotypes of all \( n \) individuals given their marker genotypes and the QTL locus (\( M \) and \( \lambda \)); and \( \pi(\lambda, \Omega, \Sigma) \) is the prior imposed on the genetic parameters. It is reasonable to assume the priors are independent for the parameters. Thus,

\[
\pi(\lambda, \Omega, \Sigma) = \pi(\lambda) \cdot \pi(\Sigma) \cdot \prod_{j=0}^{2} \pi(\Omega_j)
\]

(2.8)

Usually, there is no information available for the QTL location \( \lambda \), a uniform distribution on \([0, D_m]\) is a natural choice for it. The information about \( \Omega_j \) can be obtained relatively reliably, for which multivariate normal priors with moderate variances are used. The standard prior distribution for the inverse of the covariance matrix \( \Sigma^{-1} \) is the Wishart \((R, \rho) [10] [20]\), where the so called scale matrix \( C = R^{-1} \) represents prior structural information about \( \Sigma \) and \( \rho \) is the degree of freedom, which must be greater than \( T - 1 \) to have a proper prior. A small value of \( \rho \) gives a relative flat distribution. The Wishart prior with low degrees of freedom and a specified \( R \) is regarded as a reference (or
noninformative) proper prior. Although it has less flexibility, this distribution offers the advantage of being a conjugate prior, leading to a relative simple form for the posterior.

In the Bayesian approach, statistical inference for the unknown parameters are based on their marginal posterior distributions. Theoretically, the marginal posteriors can be obtained from the joint posterior \( \pi(\Omega, \Sigma, Q, \lambda) \) by integrating over the other unknowns. Unfortunately, in practice, the evaluation of such high-dimensional integrals in closed form is not possible. However, it is often straightforward to derive either the full conditional posterior distributions for some parameters, or the explicit expressions that are proportional to the corresponding full conditional posterior distributions for other parameters, i.e.,

\[
\pi(\Omega_j | y, \Omega_{-j}, \Sigma, Q, \lambda) \\
\propto \pi(y | \Omega, \Sigma, Q, \lambda) \cdot \pi(\Omega_j) \\
\propto \exp \left\{ -\frac{1}{2} \sum_{q_i=j} \left[ (y_i^{(j)} - g(\Omega_j))' \Sigma^{-1} (y_i^{(j)} - g(\Omega_j)) \right] - \frac{1}{2} (\Omega_j - \eta)' \Lambda^{-1} (\Omega_j - \eta) \right\} 
\]

(2–9)

and

\[
\pi(\Sigma^{-1} | y, \Omega, Q, \lambda) \\
\propto \pi(y | \Omega, \Sigma^{-1}, Q, \lambda) \cdot \pi(\Sigma^{-1}) \\
\propto |\Sigma^{-1}|^{\frac{n+p+T+1}{2}} \cdot \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} + \sum_{j=0}^2 \sum_{q_i=j}(y_i^{(j)} - g(t|\Omega_j))(y_i^{(j)} - g(t|\Omega_j))' \right] \right\} \\
\sim Wi(D^{-1}, n + \rho) 
\]

(2–10)

where \( \Omega_{-j} = \{ \Omega_{j'} : j' = 0, 1, 2, j' \neq j \} \); \( y_i^{(j)} \)'s represent for the observations from those individuals which have genotype \( j \); and \( D = R^{-1} + \sum_{j=0}^2 \sum_{q_i=j}(y_i^{(j)} - g(t|\Omega_j))(y_i^{(j)} - g(t|\Omega_j))' \).

### 2.5 Implementation

Under the Bayesian framework, with the explicit expressions described in section 2.4, a Markov chain Monte Carlo technique can be used to draw samples from the joint posterior distributions 2–7 of unknown parameters. In this work, a hybrid scheme of Gibbs sampler and Metropolis-Hastings (M-H) algorithm [29] [36] [87] will be applied.
In particular, Gibbs sampling steps update $\Sigma^{-1}$, while the M-H algorithm updates $\Omega_j$ and the QTL position. After generating a random sequence of states $(\lambda^0, Q^0, \Omega^0, \Sigma^0), (\lambda^1, Q^1, \Omega^1, \Sigma^1), ... , (\lambda^N, Q^N, \Omega^N, \Sigma^N)$, the final MCMC samples are collected, from which we are able to draw inference for the unknown parameters.

**Algorithm**: The Markov chain is constructed in the following way:

*Step 1.* Initialize the iteration at an arbitrary point $(\lambda^0, Q^0, \Omega^0, \Sigma^0)$, which has a positive posterior density.

*Step 2.* Modify the 4 blocks of unknowns parameters and move to a new state from $(\lambda^{k-1}, Q^{k-1}, \Omega^{k-1}, \Sigma^{k-1})$ through successive generation new values $\lambda^k, Q^k, \Omega^k,$ and $\Sigma^k$. More specifically, given the values of the unknowns $(\lambda, Q, \Omega, \Sigma)$ from the current state, we proceed as follows:

*Step 2-1 (updating $\lambda$)*: In each step, following the idea of Satagopan et al [78], $\lambda$ is updated by using the metropolis algorithm. A new value of $\lambda^*$ is generated from $\text{Uniform}(\max(0, \lambda - \delta), \min(\lambda + \delta, D_m))$, and this proposed distribution is denoted by $q(\lambda, \lambda^*)$. This proposed $\lambda$ is accepted with probability $\min(\alpha_\lambda, 1)$, and the state remains current value if the proposal is reject. Where $\alpha_\lambda$ is given as below:

$$\alpha_\lambda(\lambda, \lambda^*) = \frac{\pi(\lambda^* | y, Q, \Omega, \Sigma) \cdot q(\lambda^*, \lambda)}{\pi(\lambda | y, Q, \Omega, \Sigma) \cdot q(\lambda, \lambda^*)}$$  \hspace{1cm} (2–11)

Note that,

$$\pi(\lambda^* | y, Q, \Omega, \Sigma) = \pi(\lambda^* | y, Q, \Omega, \Sigma, M)$$

$$= \pi(\lambda^* | Q, M)$$

$$\propto \pi(Q | \lambda^*, M) \cdot \pi(\lambda^*)$$ \hspace{1cm} (2–12)

Similarly,

$$\pi(\lambda | y, Q, \Omega, \Sigma) \propto \prod_{i=1}^{n} \pi(Q_i | \lambda, M_i) \cdot \pi(\lambda)$$  \hspace{1cm} (2–13)
Hence, the accept probability 2–11 can be simplify to

$$\alpha_\lambda(\lambda, \lambda^*) = \min \left( \frac{\prod_{i=1}^{n} \pi(Q_i \mid \lambda^*, M_i) \cdot q(\lambda^*, \lambda)}{\prod_{i=1}^{n} \pi(Q_i \mid \lambda, M_i) \cdot q(\lambda, \lambda^*)}, 1 \right).$$

(2–14)

One of the important implementation issues here is that the tuning parameter $\delta > 0$ needs to be chosen carefully, since it determines the variance of the proposal. If the variance of the proposed density is too large, a large proportion of proposed moves will be rejected, which may result in inefficiency. On the other hand, a too small variance gives a high accept rate whereas a slow movement on the parameter space, which also leads to inefficiency. Satagopan et al [78] suggested that more than one updates of $\lambda$ between other updates may be able to improve the mixing of the Markov chain.

**Step 2-2 (updating $Q$)**: Due to the independence of $n$ individuals, $Q$ is updated via separately updating each $Q_i$. For each individual $i$ and QTL genotype $j$, the full conditional density is in the form of a multinormial with cell probabilities

$$p_{ij} = \pi(Q_i = j \mid y, \Omega, \Sigma, \lambda) = \frac{\pi(Q_i = j \mid \lambda) \cdot \pi(y \mid \Omega, \Sigma, Q_i = j)}{\sum_{q_i=0}^{2} \pi(Q_i = q_i \mid \lambda) \cdot \pi(y_i \mid \Omega, \Sigma, Q_i = q_i)},$$

(2–15)

Hence, at each cycle, we can sample the QTL genotype $Q_i$ directly from this full conditional density.

**Step 2-3 (updating $\Omega$)**: We update each $\Omega_j$ successively by a Metropolis-Hastings algorithm. For each QTL genotype, a new value $\Omega_j^*$ is generated from a proposed density $q_j(\Omega_j, \Omega_j^*)$, given current $\Omega$. Evaluate the acceptance probability of the move is $\min(1, \alpha_{\Omega_j})$. In general, $\alpha_{\Omega_j}$ can be expressed as:

$$\alpha_{\Omega_j} = \frac{\pi(\Omega_j^* \mid y, \Omega_{-j}, \Sigma, Q, \lambda) \cdot q(\Omega_j^*, \Omega_j)}{\pi(\Omega_j \mid y, \Omega_{-j}, \Sigma, Q, \lambda) \cdot q(\Omega_j, \Omega_j^*)}.$$

(2–16)

Note that, the choice of the Metropolis kernel $q$ is essentially arbitrary, and a symmetric $q$ in the sense that $q(\Omega_j, \Omega_j^*) = q(\Omega_j^*, \Omega_j)$ is usually preferred. And in that case, the ratio $q(\Omega_j^*, \Omega_j)/q(\Omega_j, \Omega_j^*)$ is canceled in the above expression 2–16. Here, for the proposed density, we use a multivariate normal distribution centered at the current $\Omega$,
with variance-covariance matrix given by an information-type matrix \([94]\) whose inverse has \((u, v)\)th element

\[
\sum_{i=1}^{n_j} \left( \frac{\partial g(t | \Omega_j)}{\partial \Omega_{j,u}} \right)' \left( \frac{\partial g(t | \Omega_j)}{\partial \Omega_{j,v}} \right) + \frac{1}{2} \frac{\partial}{\partial \Omega_{j,u}} \frac{\partial}{\partial \Omega_{j,v}} (\Omega_j - \eta)' \Sigma^{-1} (\Omega_j - \eta).
\]  

(2–17)

This expression combines the expected information (the first term) with the prior information (the second term) offers an advantage of avoiding singular information matrices. Unfortunately, tedious initial analysis has to be done to obtain estimates \(\Omega_j\) and \(\Sigma\) from which to evaluate 2–17. So, initially the Metropolis algorithm described before can be carried out by using an arbitrary variance-covariance matrix. The posterior mean of \(\Omega_j\) and \(\Sigma\) is then plugged into 2–17 from the subsequent analysis.

**Step 2-3 (updating \(\Sigma\))**: We generate a new value of \(\Sigma^{-1}\) directly from its full conditional posterior distribution. This is straightforward since it has an explicitly expression 2–10 for its full conditional posterior distribution.

### 2.6 Estimation Issues

For those parameters, such as \(\lambda\) and the curvature paramters, the corresponding Bayesian estimators are given by the empirical means of their marginal posteriors. It has been justified in Tierney \([87]\) that for any function of unknown \(f\), which is a square integrable with respect to the stationary distribution \(\pi\), if \((\lambda^{(k)}, \Omega^{(k)}, Q^{(k)}, \Sigma^{(k)})\) are the samples that we collect from the Markov Chain,

\[
\hat{f}_N = \frac{1}{N} \sum_{k=1}^{N} f(\lambda^{(k)}, \Omega^{(k)}, Q^{(k)}, \Sigma^{(k)}) \to E_\pi[f(\lambda, \Omega, Q, \Sigma | y)].
\]  

(2–18)

In other words, the empirical averages of their corresponding MCMC samples may be regarded as the consistent estimators for the unknown parameters.

For the marginal posterior densities of these parameters, kernel density estimator \([23]\) can be used, since the closed form of their full conditional posteriors are not available. For example, if the simplest kernel density estimator is utilized, then the histogram kernel
density estimator for $\lambda$ is given by

$$
\hat{\pi}(\lambda \mid y) = \frac{1}{Nh} \sum_{j=0}^{Dm/h} I(jh < \lambda \leq (j + 1)h) \sum_{k=0}^{N} I(0 < \frac{\lambda^{(k)}}{h} - j \leq 1)
$$

(2–19)

Another important estimation issue is to obtain the confidence intervals for the unknowns. Box and Tao [5] suggested that highest posterior density (HPD) regions can be constructed to give the confidence intervals for the parameters of interest and the detailed method for developing approximated HPD via MCMC samples can be seen in [76]. Alternatively, we can obtain the approximate HPD for the parameters directly by from their corresponding smooth density estimators.

Finally, in order to estimate the Monte Carlo error via the CLT, Geyer [31] suggested three types of consistent estimators, the window estimators, the method of standardized time series and the specialized Markov Chain estimators. Among those, the windows estimators probably provides the best estimates, although it requires more work and stronger regularity conditions to be consistent.

### 2.7 Modeling the Covariance Matrix by a Reference Prior

Although the Wishart is a standard prior for the covariance matrix and convenient to use, it has been criticized for being too restricted and its lack of flexibility. Also, as pointed out by Stein Dempster [15] and [84] that, in practise, if the true covariance matrix $\Sigma$ is closed to $I$, the eigenstructure of $\Sigma$ can be systematically distorted by the estimator, so this conventional prior can behave poorly, especially when the sample size is small or the data are sparse. To overcome these drawbacks, several other more flexible priors have been introduced, including a log-matrix prior [51], a reference noninformative prior [109], and a constrained Wishart prior [21]. Meanwhile, several recent papers [13] [14] [72] proposed strategies of modeling the covariance matrix with a different parameterization. All these papers are based on the key idea that a covariance matrix for longitudinal data can be diagonalized, i.e.,

$$A'\Sigma A = S,$$

(2–20)
where $S$ is a diagonal matrix with positive entries and $A$ is a unique lower triangular matrix with 1’s on the diagonal.

Without nice properties due to the conjugate priors, the resulting full conditional posterior of $\Sigma$ no longer has an explicit form. Often, one even have to generate $\Sigma$ componentwise by using Gibbs sampling. Consequently, I will focus on computationally easy methods that can generate the entire $\Sigma$ at a time, given the problem’s complexity. Among various choices, the reference (noninformative) prior was first introduced by Berger and Bernado [3] and thoroughly discussed by Yang and Berger [109]. As an option to further improve the estimation for the covariance matrix, I further investigated the implementation of this reference prior in the context of QTL functional mapping.

The most commonly used reference prior might be the Jeffrey’s prior (Jeffreys 1961),

$$\pi(\Sigma) = |\Sigma|^{-(p+1)/2}. \tag{2–21}$$

However, care must be taken when using this prior, as it can lead to an improper posterior distribution, and it might fail to shrink the eigenvalues appropriately sometimes. However, the approach proposed by Yang and Berger [109] has proven to be able to overcome these inadequacies of the Jeffrey’s prior remarkably. Note that $\Sigma$ can be decomposed as $\Sigma = OD'O'$, where $O$ is orthogonal with positive entries in the first row, and $D = \text{diag}(d_1, d_2, \ldots, d_p)$, with $d_1 \geq d_2 \geq \ldots \geq d_p$. Hence, provided these monotonically ordered $\{d_i\}$, the reference prior for $(D, O)$ is given by

$$\pi(D, O) \propto \frac{1}{|\Sigma| \prod_{i<j} (d_i - d_j)}, \tag{2–22}$$

and the resulting posterior distribution is

$$\pi(\Sigma|y, \Omega, Q, \lambda) \propto \exp[tr(-\frac{1}{2n} OD^{-1}O'(\sum_{j=0}^{2} \sum_{q=0}^{n_j} (y_{ij} - g(t|\Omega_j))(y_{ij} - g(t|\Omega_j)))'\Sigma^{-1} n/2)]^{(n/2+1)} \tag{2–23}$$

Comparing the reference prior with the Jeffrey’s prior, it is noted that the posterior given in equation 2–22 is always proper. Also note that, since this reference prior put
more mass near the region for equal eigenvalues, it can produce an estimator with better
eigenstructure.

Unsurprisingly, it is very difficult to analytically evaluate the posterior. Yang
and Berger suggested using a Metropolisized hit-and-run sampler algorithm to obtain the
integration. The detail sampling procedure at the kth iteration is given as follows:

\textit{step 1} : Given the current positive-define matrix } \Sigma_k \textit{, we set } W_k = \log \Sigma_k \textit{, in the sense that } W_k = \sum_{i=0}^{\infty} \frac{(\Sigma_k)^i}{i!}.

\textit{step 2} : Randomly generate a symmetric } p \times p \textit{ matrix } T \textit{, with elements } t_{ij} = z_{ij}/(\sum_{l \leq m} z_{lm}^2)^{1/2}, \textit{where } z_{ij} \sim i.i.d.N(0,1), \textit{for } i \leq j.

\textit{step 3} : Set } W^* = W_k + \nu T \textit{ where } \nu \textit{ is generated from } N(0,1).

\textit{step 4} : Update } W_k \textit{ with an accept probability } \min(1, \pi^*(W^*|y, \Omega, Q, \lambda)/\pi^*(\Sigma_k|y, \Omega, Q, \lambda)).

2.8 Modeling the Structure of the Covariance Matrix

As to the longitudinal data analysis, parametric covariance modeling has several
advantages comparing to the conventional multivariate approach, which ignores the
covariance structure. First, probably the biggest advantage is that parametric modeling
covariance structures generally results in more efficient estimation of covariance matrix
and hence the mean structure. Second, it can be used to deal with data more efficiently,
when the number of measurement times is relatively large. And finally, it can be used to
handle the data more effectively when time-points of the measurements are not common
across all subjects. For those reasons, during the last tow decades, it became an important
topic to develop explicit parametric models for the data’s variance-covariance matrix
structure.

2.8.1 Modeling the Structure of the Within-Subject Covariance Matrix by
the First-Order Autoregressive Model

One common covariance structure exhibited by longitudinal data is serial correlation.
For example, Table 2-3 displays the matrix of sample variance and correlations corresponding
the real data for the body mass of mice measured at 10 consecutive weeks. Clearly, the
variances are not homogeneous, and tend to increase over time. In order to stabilized the
variance, a log-transformation is used. Unfortunately, as in our case, data transformations can bring us troubles to interpret the curvature parameters biological manfully. To handle this difficulty, we incorporate the so called transform-both-sides (TBS) method \[9\] into our model to further preserve the biological means for these curvature parameters. Table 2-2 displays the sample variance-covariance matrix for the transformed data. Notice that, first, after the transformation, the variances tend to be a constant over time. Second, the correlations exist apparently, and they are mostly positive. For any given column, correlations tend to decrease to zero. Finally, the serial correlation seems to be present, since the correlations decrease as the elapsed time between measurements increases.

Stationary autoregressive (AR) models \[43\] \[44\] are perhaps the most popular parametric models for serial correlations. These models are based on two assumptions, variances are constant over time and correlations between measurements with equal time lag are equal. Particularly, in our problem, the autoregressive model with order 1, i.e. the AR(1) model is considered. As a result, the covariance matrix $\Sigma$ for observations $y_i = (y_{i1}, \ldots, y_{iT})$ can be constructed as follows

$$cov(y_{t_j}, y_{t_j}) = \sigma^2; \quad \forall 1 \leq t_j \leq T$$

(2–24)

and

$$cov(y_{t_j}, y_{t_k}) = \sigma^2 \cdot \rho^{|t_j - t_k|}; \quad \forall 1 \leq t_j \neq t_k \leq T.$$  

(2–25)

As a special case of parametric model, Bayesian analysis can be performed to obtain the estimates of these matrix structure parameters. Again, it is reasonable to assume the independency between the priors of $\sigma^2$ and $\rho$. As a conventional choice, the inverse gamma prior is given to $\sigma^2$, and an informative prior restricted on $[-1, 1]$ is imposed on $\rho$. In such a case, the posterior density of $\lambda, \Omega, \sigma^2, \rho$, and $Q$ is given by:

$$\pi(\lambda, \Omega, Q, \sigma^2, \rho \mid y) = \pi(y \mid \Omega, Q, \sigma^2, \rho) \cdot \pi(Q \mid \lambda) \cdot \pi(\lambda) \cdot \pi(\sigma^2) \cdot \pi(\rho) \cdot \prod_{j=0}^{2} \pi(\Omega_j), \quad (2–26)$$

where $\pi(\sigma^2) = IG(\alpha, \beta)$ and $\pi(\rho) = Uniform(-1, 1)$.  

36
The closed form of the full conditional posterior distributions of $\sigma^2$ and $\rho$ are not available, just as some of the other parameters, but the explicit expressions that are proportional to their corresponding full conditional posteriors can be derived. For $\sigma^2$,

$$
\pi(\sigma^2 \mid y, \lambda, Q, \Omega, \rho) \propto \pi(y \mid \lambda, Q, \Omega, \sigma^2, \rho) \cdot \pi(\sigma^2)
$$

$$
\propto |\Sigma(\sigma^2, \rho)|^{-\frac{3}{2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \sigma^{2(-\alpha-1)}
$$

$$
\cdot \exp\left\{-\beta \frac{\sigma^2}{\rho} - \frac{1}{2} \sum_{j=0}^{2} \sum_{q_{ij}=j}^n (y_i - f(\Omega_j))' \Sigma^{-1}(\sigma^2, \rho)(y_i - f(\Omega_j))\right\}
$$

(2–27)

and for $\rho$,

$$
\pi(\rho \mid y, \lambda, Q, \Omega, \sigma^2) \propto \pi(y \mid \lambda, Q, \Omega, \sigma^2, \rho) \cdot \pi(\rho)
$$

$$
= \frac{1}{2} \cdot |\Sigma(\sigma^2, \rho)|^{-\frac{3}{2}} \cdot \exp\left\{-\frac{1}{2} \sum_{j=0}^{2} \sum_{q_{ij}=j}^n (y_i - f(\Omega_j))' \Sigma^{-1}(\sigma^2, \rho)(y_i - f(\Omega_j))\right\}
$$

(2–28)

Based on above expressions, the corresponding Metropolis-Hastings steps can be developed to update $\sigma^2$ and $\rho$ within the MCMC estimation scheme described in section 2.5.

**updating $\sigma^2$ :**

In each MCMC cycle, a candidate value of $\sigma^2$ denoted by $\sigma^2*$ is generated from its proposal distribution, which can be specified as: $q(\sigma^2^* \mid \sigma^2) = IG(\frac{1}{\sigma^2}, 1)$. And this proposal will be accepted with probability $\min(\alpha_{\sigma^2}, 1)$, where

$$
\alpha_{\sigma^2} = \frac{\pi(\sigma^2^* \mid y, \lambda, Q, \Omega, \rho) \cdot q(\sigma^2^* \mid \sigma^2)}{\pi(\sigma^2 \mid y, \lambda, Q, \Omega, \rho) \cdot q(\sigma^2 \mid \sigma^2^*)}.
$$

(2–29)

**updating $\rho$ :**

The proposal distribution of $\rho$ can be specified as a uniform with in moderate range around the current value of $\rho$, in other words, $q(\rho^* \mid \rho) = Uniform(max(-1, \rho - \delta_\rho), min(\rho + \delta_\rho, 1))$. A new value of $\rho$, $\rho^*$, is generated from this proposal distribution and is accepted with probability $\min(\alpha_{\rho}, 1)$, with

$$
\alpha_{\rho} = \frac{\pi(\rho^* \mid y, \lambda, Q, \Omega, \sigma^2) \cdot q(\rho^* \mid \rho)}{\pi(\rho \mid y, \lambda, Q, \Omega, \sigma^2) \cdot q(\rho \mid \rho^*)}.
$$

(2–30)

37
2.8.2 Modeling the Structure of the Within-Subject Covariance Matrix by the Structured Antedependence Model

Often, efforts to stabilize variances by transformations meet only limited success, and we need to loose both restrictions on the constant variances and equal correlations between measurements that are equidistant in time. Considering observations for one individual $y_1, y_2, \ldots, y_T$, whose joint distribution is multivariate normal. These observations are called $r$-order antedependent \([27]\) if $y_t$ given $y_{t-1}, \ldots, y_2, y_1$ only depends on $y_{t-1}, \ldots, y_{t-r}$, for all $t \geq r$. Based on the assumption that observations within individuals are antedependent, Núñez-Antón and Zimmerman \([63]\) proposed a useful class of models called structured antedependence (SAD) models. The idea of this approach is to fit the covariance structure by parametric modeling the antedependence coefficients and innovation variances. Hence as a result, the SAD model has a great advantage of reducing the number of parameters considerably compared with the traditional unstructured antedependence (UAD) models.

To be specific, if an $r$-order SAD model is assumed, the autoregressive coefficients follow a Box-Cox power law and the innovation variances are specified as parsimonious functions of time, i.e.,

$$
\phi_{j,k} = \phi_k^{f(t_j; \lambda_k) - f(t_{j-k}; \lambda_k)}, \quad j = r + 1, \ldots, T; k = 1, \ldots, r
$$

$$
\nu_j^2 = \nu^2 \cdot g(t_j; \psi) \quad j = r + 1, \ldots,
$$

where

$$
f(t; \lambda) = \begin{cases} 
\frac{(t^{\lambda-1})}{\lambda} & \text{if } \lambda \neq 0 \\
\log(t) & \text{if } \lambda = 0, 
\end{cases}
$$

and here $g(\cdot)$ is a relatively simple function with fewer parameter, say a polynomial with low-order.
Considering the complexity of our particular problem, a simplified SAD(1) model with constant innovation variances is incorporated to the QTL functional mapping of longitudinal growth traits. The analytical forms for variance and covariance functions among equally spaced time-dependent measurements, were given by Jaffrézic et al.\[41\],

\[
\sigma^2(j) = \frac{1 - \phi^{2j}}{1 - \phi^2} \nu^2, \quad \sigma^2(j, k) = \phi^{t_j - t_k} \frac{1 - \phi^{2k}}{1 - \phi^2} \nu^2 \quad 1 \leq t_k \leq t_j \leq T.
\]

Therefore, different from an AR(1) model, even for the simplest SAD model which has constant innovation variances over time, neither the variance of the observed process nor the correlation function is stationary, i.e. the variances \(\sigma^2(j)\) can change with time, and \(\sigma^2(j, k)\) does not depend only on the lag time \(|j - k|\).

Within the framework of Bayesian analysis, a thick-tailed inverse-gamma prior can be given to the innovation variance \(\nu^2\), and a normal prior can be given to the antedependence parameter \(\phi\). In the actual analysis, the priors are selected as, \(\pi(\nu^2) = IG(\alpha, \beta) = IG(1, 1)\) and \(\pi(\phi) = N(\mu_\phi, \eta_\phi) = N(0, 10)\). As before, for both parameters, the explicit expressions that are proportional to their corresponding full conditional posteriors can be derived as follows,

\[
\pi(\nu^2 | y, \lambda, Q, \Omega, \phi) \propto \pi(y | \lambda, Q, \Omega, \nu^2, \phi) \cdot \pi(\nu^2)
\]

\[
\propto |\Sigma(\nu^2, \phi)|^{-\frac{n}{2}} \cdot \frac{\beta^n}{\Gamma(n)} \cdot \nu^{2(-\alpha - 1)} \cdot \exp\left\{-\frac{\beta}{\nu^2} - \frac{1}{2} \sum_{j=0}^{2} \sum_{q_i=j}^{n_j} (y_i - f(\Omega_j))' \Sigma^{-1}(\nu^2, \phi)(y_i - f(\Omega_j))\right\}
\]

and for \(\phi\),

\[
\pi(\phi | y, \lambda, Q, \Omega, \nu^2) \propto \pi(y | \lambda, Q, \Omega, \nu^2, \phi) \cdot \pi(\phi)
\]

\[
= \frac{1}{2} \cdot |\Sigma(\nu^2, \phi)|^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2\eta_\phi}(\phi - \mu_\phi)^2 - \frac{1}{2} \sum_{j=0}^{2} \sum_{q_i=j}^{n_j} (y_i - f(\Omega_j))' \Sigma^{-1}(\nu^2, \phi)(y_i - f(\Omega_j))\right\}.
\]

As long as we specified the proposal distributions, the detailed Metropolis-Hastings steps updating \(\nu^2\) and \(\phi\) can be derived as below,
updating $\nu^2$:

In each MCMC cycle, a candidate value of $\nu^2$ denoted by $\nu^{2*}$ is generated from its proposal distribution, which can be specified as: $q(\nu^{2*} | \nu^2) = IG(\frac{1}{\nu^2} + 1, 1)$. And this proposal will be accepted with probability $min(\alpha_{\nu^2}, 1)$, where

$$\alpha_{\nu^2} = \frac{\pi(\nu^{2*} | y, \lambda, Q, \Omega, \phi) \cdot q(\nu^{2*} | \nu^2)}{\pi(\nu^2 | y, \lambda, Q, \Omega, \phi) \cdot q(\nu^2 | \nu^{2*})}.$$  \hspace{1cm} (2–36)

updating $\phi$:

The proposal distribution of $\rho$ can be specified as a uniform with in moderate range around the current value of $\rho$, in other words, $q(\phi^* | \phi) = N(\phi, V_\phi)$. A new value of $\phi$, $\phi^*$, is then generated from this proposal distribution and is accepted with probability $min(\alpha_\phi, 1)$, with

$$\alpha_\phi = \frac{\pi(\phi^* | y, \lambda, Q, \Omega, \nu^2) \cdot q(\phi^* | \phi)}{\pi(\phi | y, \lambda, Q, \Omega, \nu^2) \cdot q(\phi | \phi^*)}. $$  \hspace{1cm} (2–37)

2.9 A Worked Example

2.9.1 Animal Material

Vaughn et al. [90] constructed a linkage map with 96 microsatellite markers for 1043 F$_2$ mice (503 males and 540 females) derived from two strains, the Large (LG/J) and Small (SM/J). The total length of this map is $\sim 1780$ cM (in Haldane’s units) and an average marker interval length is $\sim 23$ cM. The F$_2$ progeny was measured for their body mass at 10 weekly intervals starting at age 7 days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth, parity and sex [90]. Overall, about 10% of the marker genotypes were randomly missing. The mice with missing data were excluded from the analyses.

2.9.2 Results

Zhao et al [113] and Zhao et al [114] first analyzed this data set by using functional mapping with maximum-likelihood based methods. They showed that body masses in the F$_2$ mice follow a logistic curve, but display substantial variation in the shapes of
curves. Bayesian-based functional mapping was used to genome-wide search for the QTL that control growth curves in mice by estimating genotype-specific curve parameters ($\Omega$), the QTL locus $\lambda$, and the covariance matrix $\Sigma$. The Bayesian formulation requires specifying prior distributions for these parameters. According to several related studies, some information was gathered for $\Omega$, and the priors for $\Omega_j$, ($j = 1, 2, ..., 3^s$) were given by a multivariate normal, centered at $(30, 10, 0.6)^T$ and with a large dispersion $\text{diag}(9, 4, 1)$. The prior of the QTL location, $\lambda$, was simply chosen to be a uniform distribution along the entire linkage group, i.e. $[0, D_m]$. The prior for the covariance matrix was set to be inverse-Wishart($R^{-1}, \rho$), where $R$ was given by the sample covariance matrix.

As suggested by Geman and Geman [29], one long run may be more efficient with considerations of using the following two strategies: (1) discard a number of initial "burn-in" simulations, and consider only the remaining samples, since it would be unlikely for those initial simulations came from the stationary distribution targeted by the Markov chain; (2) subsampling the chain at equal intervals to reduce the serial correlation among samples. Also, based on the detailed discussions in Gelfand and Smith [30], several graphical techniques can be applied here to check the convergence. Initially, as pilot runs, several short Markov Chains are started at different QTL locations along a linkage group. It appears that all of these short chains converge to a relatively same stationary distribution after roughly 10,000 cycles. Also, for these test runs, the autocorrelation function of the single locus are sampled at different spans; as displayed in figure 2-1. The estimated autocorrelation function of MCMC samples at lag $k$ can be expressed as

$$ \hat{R}(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} [x_t - \mu] \cdot [x_{t+k} - \mu]. \quad (2-38) $$

We observed from these plots that autocorrelation between the every 60th sub-sample decreased quickly, and is closed to 0 at lag 10.

According to the results from these pilot runs, the starting value for the single putative locus was set to be $\lambda_0 = 40$ cM for all linkage groups. For each analysis, the
Markov Chain ran 60,000 cycles and sampled every 60 cycles, which yielded a working set with 1000 states. These states were regarded as samples from the targeted posterior distributions of the unknowns.

An informal check of parameter identifiability is conducted via on graphical techniques based on time series theories [17]. As shown in Figure 2-2, the consecutive samples move randomly towards different directions, which indicates that the MCMC sampler is not "sticky". This means that the parameters can be regarded as being identifiable.

Estimated marginal posteriors of QTL location for all 19 chromosomes are displayed in figure 2-3. At a first glance, we noticed that posteriors on chromosome 6, 7, 8, 10, 11, and 15 have obvious spikes, whereas those on other chromosomes display relatively flatter
patterns along the whole chromosome. By comparing the Bayes factors of $M_0$ and $M_1$, single QTL’s are detected on chromosome 6, 7 and 10 respectively. Where the logarithmic scaled Bayes factors computed by $3-5$ is 12.91 for chromosome 6, 13.47 for chromosome 7, and 7.99 for chromosome 10 respectively. A single QTL was estimated on 82.7 cM between marker 3 and 4 of chromosome 6, or on 46.8 cM between marker 2 and 3 of chromosome 7, or on 77.7$cM$ between marker 3 and 4 of chromosome 10. These results are generally consistent with those in Zhao et al[114]. It is straightforward to make inference on the unknown parameters based on these marginal posteriors. Table 2-4, 2-5, and 2-6, show the summarized results from the analysis in which the estimated parameters are given by their
corresponding posterior means. The numbers in the parentheses are the 95% equal-tail confidence intervals, which were obtained as HPD regions.

Figure 2-3. A profile of Estimated marginal posterior distribution of the QTL location by assuming that exactly one QTL is located on one of the chromosome respectively.

Estimated curve parameters are listed in Table 2-4, 2-5, and 2-6, assuming the actual QTL is located on chromosome 6, 7, or 10. And given a set of estimated curve parameters, we can obtain the growth curves by using the logistic growth function 2–2. Figure 2-4, 2-5, and 2-6 illustrate the growth curves of the three genotypes at each of the detected QTL. Based on quantitative genetic theory, we can partition time-dependent genotypic value, \( \mu_j(t) (j = 0, 1, 2) \), into the additive and dominant effects due to the QTL. To be specific, the dynamic changes of additive genetic effects of the QTL can be expressed as

\[
a(t) = \frac{1}{2} [\mu_2(t) - \mu_1(t)],
\]  

(2–39)

and the dynamic changes of dominant genetic effects of the QTL can be expressed as

\[
d(t) = \frac{1}{2} [2\mu_1(t) - \mu_2(t) - \mu_0(t)].
\]  

(2–40)
Estimated by above equations, dynamic patterns of both additive and dominant effects were plotted in figure 2-7. The estimated additive effect is positive and increases with age for all three detected QTLs. On the other hand, the estimated dominant effect displays different patterns depending on the QTL detected. And also note that, the magnitude of dominant effect is much smaller comparing to the additive effects.

Figure 2-4. Fitted growth curves for the three QTL genotypes assuming a single QTL is located on mouse chromosome 6.

By fitting single-QTL model, I detect 3 QTL on chromosome 6, 7, and 10 separately, and this fact suggests that there might actually exist more than one QTL controlling the growth curves of mice. To further explore this possibility, I also fit the data with a multi-QTL model, which will be described in the next chapter.
2.10 Monte Carlo Simulation

In this section, several simulated numerical examples are presented to illustrate the introduced Bayesian approach. Most importantly, I would like to compare the results produced by using our new approach to those results produced by using the conventional maximum-likelihood based functional mapping method. Besides that, I would also like to see how the estimates will be influenced when using different strategies to estimate the covariance matrix. The designs of these simulation studies were inspired by the actual mice-body-mass experiment that I described earlier.

Simulation experiments were performed to investigate the statistical properties of this Bayesian model proposed for functional mapping for complex dynamic traits. The experiment was designed as follows: An F$_2$ population of 450 individuals was simulated,
and I assume that there is exactly one QTL located on a linkage group with 11 equally scattered markers. The total length of this linkage group is 100 cM, and I assume that there is exactly one QTL, and it is located at 34 cM. The true values of the curvature parameters were chose to be:

$$\Omega_{QQ} = (36.7, 11.9, 0.65); \quad \Omega_{Qq} = (35.6, 11.2, 0.64); \quad \Omega_{qq} = (33.4, 11.2, 0.65).$$

The sample covariance from the dog-body-mass data was set to be the true covariance matrix. Given this setting of covariance matrix and the curvature parameters of the 3 genotypes, the heritability of the QTL is about 0.1. Finally, vectors of phenotypic values are simulated at 10 evenly spaced time points for these 450 F$_2$ individuals.
Figure 2-7. Dynamic changes of the additive and dominant effect due to the QTL located on mouse chromosome 6, 7, and 10 respectively.

In order to implement the MCMC techniques described in section 2.5, suitable priors have to be chosen. The curve parameters $\Omega_j$ is given a multivariate normal distribution centered at $(30, 10, 0.7)$ with covariance matrix $\Lambda = \text{diag}\{10, 5, 4\}$. Since we don’t have any prior information about the QTL location, $\lambda$ is simply assumed to have uniform prior along the entire linkage group. Three methods were used to estimate the covariance matrix of the phenotypical values. As the conventional choice, method 1 imposed an inverse Wishart prior with degree of freedom $T = 10$ on $\Sigma$. Based on section 2.7.2, method 2 utilized the SAD(1) to model the covariance structure. I imposed $IG(1, 1)$ as a prior on the parameter of innovation variance $\nu^2$, and $Normal(0, 10)$ as a prior on the antedependence parameter $\phi$, respectively. Finally, method 3 modeled the covariance structure by the AR(1) model. Based on what we described in section 2.7.1, We imposed $IG(10, 1)$ as a prior on the variance $\sigma^2$, and $Uniform(-1, 1)$ as a prior on the parameter of correlation $\rho$, respectively.
Figure 2-8. Estimated marginal posterior distribution of the QTL location on mouse chromosome 6, 7, and 10 respectively.

For each MCMC-implemented Bayesian analysis, 10,000 initial "burn-in" iterations were discard, and for the rest 60,000 cycles, samples are collected for every 60 cycles, which result in a working set of 1000 states. Again, in order to ensure the chain had converged to the same stationary distribution, I repeat the MCMC experiment several times with the same simulated phenotypical data set. Those replicates produced results that were generally consistent for each method. Hence, for the purpose of demonstration, I present one of the replicates to show the general behavior of these 3 different methods.

Estimates yielded by using method 1 method 3 are displayed in table 2-7, 2-8, and 2-9 respectively. The true location of QTL and the true values of the parameters that defining the growth curve are given in square brackets, also the 95% empirical HPD (highest posterior density) confidence regions are given in the parentheses.
It can be seen from the results that among three methods, method 1 (unstructured covariance matrix) and method 2 (model the covariance matrix by SAD(1)) provided much better estimates than method 3 (model the covariance matrix by AR(1)). Both method 1 and method 3 provide a good fit to the simulated data, with reasonably small biases and narrow confidence intervals. Method 1 had better precision estimating the curvature parameters, while method 3 provided slightly narrower confidence interval for the QTL location $\lambda$ (as displayed in figure 2-9, 2-10, and 2-11). As to the computational load, method 1 had a great advantage to require only $\frac{3}{4}$ of the computational time of method 3.

Figure 2-9. Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming no covariance structures.

For the purpose of comparison, I also analyzed the same simulated data by using the conventional maximum-likelihood based method. Standard test procedures were performed to determine the location for a QTL. The Likelihood ratio test statistic was computed for
Figure 2-10. Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming the covariance structure is SAD(1).

every 2 cM along the whole linkage group. Figure 2-12 shows the profile of these values of Likelihood ratio test statistic. This resulting profile had two peaks. A larger peak with LR score 635.71 was located at 34cM, and a relatively smaller peak with LR score 628.91 was located at 44cM. Therefore, the traditional method was also able to detect the location of the simulated QTL accurately, but with a high risk of detecting false QTLs. Although the estimates of curvature parameters (see Table 2-10) were generally resalable, these estimates had significantly bigger biases than those given by the new method we proposed. Most importantly, by a single scan, the traditional method is incapable of producing confidence intervals of the estimates, or determine the critical value of the LR test. A permutation test can be performed to determine the critical value, but requires extremely
Figure 2-11. Estimated marginal posterior distribution of the QTL location on a mouse chromosome in a simulation study, assuming the covariance structure is AR(1).

extensive computational time which can cost 20 times of the computational time by using our new approach.

To further investigate the behaviors of these estimators, more replicates of this simulation experiment need to be conducted. And for every analysis, although I didn’t encounter any complications, we need to always keep it in mind that a sensitivity analysis is necessary to examine the dependence of the parameter estimates on the priors and initial values. Under the same MCMC sampling scheme, we can initiate the chain with several different starting point and/or with different choices of priors [74]. If the estimated parameters are not sensitive to the priors and the initial values, those runs should exhibit similar and stable behavior.
2.11 Discussion

The genetic architecture of complex traits can be well understood by incorporating their developmental features described by mathematical functions. Functional mapping that integrates genetics, statistics and developmental biology can be useful for deciphering the ontogenetic development of the genetic control of a complex trait [107]. Original models for functional mapping were derived within the maximum likelihood (ML) context and implemented with the EM algorithm. Although ML-based approaches possess many favorable statistical properties in parameter estimation, they may not be powerful enough to handle the complexity of high-dimensional QTL mapping models, as often seen in functional mapping. As an increasingly popular approach, Bayesian methods display
remarkable capacity to estimate genetic parameters in QTL mapping [39] [40], and [86] [88] [78] [81] [111].

In this chapter, I derived a general Bayesian framework for functional QTL mapping of dynamic traits and implemented Markov chain Monte Carlo (MCMC) algorithms to locate genomic positions of QTLs, and estimate the mathematical parameters that define a biological process and the statistical parameters that model the covariance structure. The Bayesian-based model allows the estimation of these parameters and their confidence intervals based on posterior distributions, and has great power to handle complex estimation issues related to functional mapping in an effective way. Like original parametric functional mapping [54], the new model allows to approximate the ontogenetic changes of the genetic effects triggered by a QTL. Because many biological processes, such as growth, follow a particular pattern of development [99], the ontogenetic control of a QTL can be mathematically described and, thereby, tested by estimating the parameters that define a biological process. The new model also take an advantage of functional mapping to model the structure of covariance matrix by a stationary or nonstationary approach. Because of its computational simplicity, the AR(1) is advantageous for structuring the covariance although it needs the variance and covariance stationarity assumptions. The TBS-based model can relax the assumption of variance stationarity [106], but it has not resolved the covariance stationarity issue when embedded into the AR(1) model. A so-called structured antedependence (SAD) model, advocated by Zimmerman and Núñez-Antón [117], can be used to simultaneously model the time-dependent changes of and variance and correlation in the analysis of longitudinal traits. The SAD model is found to display many favorable properties [117].

The use of Bayesian approaches to functional mapping was also considered by other authors. Yang and Xu [110] integrated Bayesian shrinkage approaches to map dynamic QTL, but their model was based on a nonparametric Legendre polynomial fitting. Although this treatment may be statistically flexible, its biological relevance may
be limited because no biologically sensible functions are deployed. The model was used to reanalyze a published data set on the growth of body mass in mice, confirming the discovery of a few QTL detected by conventional ML-based functional mapping [113] [114]. Yet, the new model provides estimates of confidence intervals of curve parameters, thus allowing better statistical inferences about the genetic control of dynamic QTLs. Simulation studies show that the new model is robust in that it provides reasonable estimates of QTL effects and positions in a wide range of parameter space.

The model proposed in this chapter can be modified by considering a network of genetic control. As a basic framework, the one-QTL is not adequate to explore the effects of interaction between different QTL [46] and QTL and environments [113] on variation in a dynamic trait. One of the significant advantages of Bayesian approaches lies in the estimation of the optimal number of QTLs involved. Variable selection via stepwise regression is used in ML mapping [45], but it is highly computationally expensive. Corresponding to this variable selection procedure, reversible jump MCMC is proposed in Bayesian analysis [33] [34], although it is subject to poor mixing and a slow convergence to the stationary distribution ([6] [32] [34]. More efficient methods based on Bayesian shrinkage analysis [108] [96] and stochastic search variable selection [112] have now been proposed. These methods do not rely upon any explicit form of variable selection; rather they proceed implicitly by shrinking the effects of excessive QTLs to zero. The modified model will certainly prove its value in elucidating the genetic architecture of dynamic traits and will probably be the beginning of detecting the driving forces behind dynamic genetics and its relationship to the organism as a whole.
Table 2-1. The sample variance-covariance matrix for the mice-body-mass data.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.420</td>
<td>0.507</td>
<td>0.683</td>
<td>0.889</td>
<td>0.801</td>
<td>0.753</td>
<td>0.791</td>
<td>0.765</td>
<td>0.756</td>
<td>0.767</td>
</tr>
<tr>
<td>2</td>
<td>0.507</td>
<td>1.060</td>
<td>1.288</td>
<td>1.477</td>
<td>1.370</td>
<td>1.469</td>
<td>1.470</td>
<td>1.475</td>
<td>1.524</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.683</td>
<td>1.288</td>
<td>2.292</td>
<td>2.591</td>
<td>2.500</td>
<td>2.327</td>
<td>2.481</td>
<td>2.461</td>
<td>2.555</td>
<td>2.570</td>
</tr>
<tr>
<td>5</td>
<td>0.801</td>
<td>1.477</td>
<td>2.500</td>
<td>4.414</td>
<td>5.467</td>
<td>5.147</td>
<td>5.884</td>
<td>6.200</td>
<td>6.371</td>
<td>6.587</td>
</tr>
<tr>
<td>6</td>
<td>0.753</td>
<td>1.370</td>
<td>2.327</td>
<td>4.184</td>
<td>5.417</td>
<td>6.491</td>
<td>7.090</td>
<td>7.593</td>
<td>7.866</td>
<td>8.231</td>
</tr>
</tbody>
</table>

Table 2-2. The sample variance-covariance matrix for the transformed mice-body-mass data.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.019</td>
<td>0.014</td>
<td>0.012</td>
<td>0.011</td>
<td>0.007</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.014</td>
<td>0.017</td>
<td>0.014</td>
<td>0.011</td>
<td>0.008</td>
<td>0.007</td>
<td>0.007</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>3</td>
<td>0.012</td>
<td>0.014</td>
<td>0.016</td>
<td>0.012</td>
<td>0.009</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>4</td>
<td>0.011</td>
<td>0.011</td>
<td>0.012</td>
<td>0.014</td>
<td>0.010</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>0.007</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>6</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.009</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>7</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>8</td>
<td>0.006</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
<td>0.011</td>
</tr>
<tr>
<td>9</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.012</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>10</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.011</td>
<td>0.012</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Table 2-3. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.09 (35.20, 37.04)</td>
<td>34.94 (34.36, 35.52)</td>
<td>33.12 (32.36, 33.93)</td>
</tr>
<tr>
<td>β</td>
<td>11.93 (11.44, 12.45)</td>
<td>11.58 (11.16, 12.03)</td>
<td>11.07 (10.65, 11.51)</td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.65 (0.64, 0.67)</td>
</tr>
<tr>
<td>QTL location</td>
<td>82.68 (67.77, 92.96)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2-4. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.09 (35.20, 37.04)</td>
<td>34.94 (34.36, 35.52)</td>
<td>33.12 (32.36, 33.93)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.93 (11.44, 12.45)</td>
<td>11.58 (11.16, 12.03)</td>
<td>11.07 (10.65, 11.51)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.65 (0.64, 0.67)</td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>82.68 (67.77, 92.96)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-5. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 7. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.55 (35.50, 37.73)</td>
<td>35.61 (34.56, 36.50)</td>
<td>33.38 (32.54, 34.33)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.83 (11.43, 12.34)</td>
<td>11.27 (10.90, 11.73)</td>
<td>11.25 (10.76, 11.70)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.63, 0.66)</td>
<td>0.64 (0.63, 0.65)</td>
<td>0.65 (0.63, 0.66)</td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>46.84 (38.80, 56.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-6. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>35.41 (34.33, 36.52)</td>
<td>34.71 (33.67, 35.70)</td>
<td>33.59 (32.61, 34.42)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.67 (11.23, 12.19)</td>
<td>11.47 (11.23, 11.81)</td>
<td>11.01 (10.61, 11.44)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.64 (0.63, 0.66)</td>
<td>0.65 (0.63, 0.66)</td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>77.78 (68.75, 80.96)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-7. Results from a simulation study by assuming none covariance structure. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>Qq</th>
<th>qq</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.67 (35.81, 37.46)</td>
<td>36.03 (35.47, 36.63)</td>
<td>33.67 (32.67, 34.31)</td>
<td></td>
</tr>
<tr>
<td>True α</td>
<td>36.70</td>
<td>35.60</td>
<td>33.40</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.83 (11.95, 12.64)</td>
<td>11.22 (10.97, 11.50)</td>
<td>11.30 (10.94, 11.60)</td>
<td></td>
</tr>
<tr>
<td>True β</td>
<td>11.90</td>
<td>11.20</td>
<td>11.20</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.66 (0.64, 0.67)</td>
<td>0.64 (0.63, 0.65)</td>
<td>0.64 (0.63, 0.66)</td>
<td></td>
</tr>
<tr>
<td>True γ</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>32.74 (28.02, 38.16)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True QTL locus</td>
<td>34.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2-8. Results from a simulation study by assuming the covariance structure to be SAD(1). Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.58 (35.85, 37.36)</td>
<td>35.88 (35.37, 36.39)</td>
<td>33.81 (33.09, 34.55)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True α</td>
<td>36.70</td>
<td>35.60</td>
<td>33.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>12.04 (11.65, 12.45)</td>
<td>11.27 (11.01, 11.54)</td>
<td>11.46 (11.09, 11.83)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True β</td>
<td>11.90</td>
<td>11.20</td>
<td>11.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.66 (0.65, 0.67)</td>
<td>0.64 (0.63, 0.65)</td>
<td>0.65 (0.64, 0.65)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True γ</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>34.63 (33.01, 36.30)</td>
<td>34.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-9. Results from a simulation study by assuming the covariance structure to be AR(1). Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>36.60 (35.59, 37.66)</td>
<td>35.57 (34.88, 36.35)</td>
<td>33.61 (32.65, 34.54)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True α</td>
<td>36.70</td>
<td>35.60</td>
<td>33.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>12.04 (11.61, 12.48)</td>
<td>11.23 (10.99, 11.56)</td>
<td>11.42 (11.05, 11.83)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True β</td>
<td>11.9</td>
<td>11.20</td>
<td>11.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.64 (0.63, 0.65)</td>
<td>0.66 (0.64, 0.67)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True γ</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>33.54 (25.54, 41.39)</td>
<td>34.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-10. Results from a simulation study by performing traditional maximum-likelihood typed method. Numbers in the box bracts are the given values of the parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
<th>QQ</th>
<th>QTL genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>37.22 [36.70]</td>
<td>37.76 [35.60]</td>
<td>33.38 [33.40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.66 [0.65]</td>
<td>0.63 [0.64]</td>
<td>0.64 [0.65]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL locus</td>
<td>34.00 [34.00]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 3
A BAYESIAN MODEL FOR MAPPING EPISTATIC QTLS THAT REGULATE
DYNAMIC PROCESSES

3.1 Introduction

Trait development is a complex process that involves multiple stages in each of which
a network of genes interact with each other and with various environments to determine
a final phenotype [25] [59]. Thus, to better understand the biology of trait development,
a detailed genetic architecture of how genes act and interact to control various stages of
development must be quantified. Unfortunately, many current statistical techniques used
in genetic research assume the additive control of genes, aimed to facilitate data analysis
and modeling, which certainly provide misleading results when genetic interactions or
epistasis actually occur.

The genetic effect or variance of quantitative trait loci (QTL) includes two components,
additive, due to the cumulation of breeding values, and nonadditive, due to allelic
(dominant) or nonallelic (epistatic) interactions. Epistatic interactions between different
loci can be further partitioned into different types, additive × additive, additive ×
dominant (or dominant × additive) and dominant × dominant. The presence of
epistasis implies that the influence of a gene on the phenotype depends critically upon
the context provided by other genes. In the past, the estimation of the additive and
nonadditive genetic architecture of a quantitative trait was based on the phenotypes
of related individuals [53], although this has minimal power to detect the nonadditive
genetic variances, especially epistatic variance because epistasis contributes little to the

The advent of DNA-based linkage maps opens a novel avenue for precisely estimating
the genetic architecture of developmental traits [90]. Current statistical methods proposed
to detect the main and interaction effects of QTL are based on the phenotypes of a
quantitative trait measured at a limited set of landmark ages. More recently, Wu et al.
[104] [105] [106] [107] and Ma et al. [54] have derived a powerful functional mapping
method for estimating the dynamic changes of QTL effects during a course of ontogenetic growth through the implementation of universal growth laws [99] and the structured residual (co)variance matrix among different time points (see [48] [71]). This method has proven to be statistically more powerful and more precise because of a reduced number of parameters being estimated, and to be biologically more meaningful due to the consideration of biological principles underlying trait development [106]. However, this model has not incorporated the estimation process of epistatic interactions and, thus, cannot examine the role of the entire genetic architecture in developmental trajectories.

In this chapter, I extend the functional mapping method to map any QTL (including additive, dominant and epistatic) that transforms allelic and/or nonallelic effects into final phenotypes during a continuous process of development represented as ontogenetic trajectories or a path through phenotype-time space [1] [100]. I derive special procedures to estimate and test the impact of epistasis on trait growth because a growing body of evidence now shows that epistasis plays a more important role in determining developmental changes than originally thought [75] [100]. I use a Bayesian-based method, implemented with the Markov chain Monte Carlo (MCMC) algorithm, to estimate QTL locations and genetic effects on growth differentiation. Compared with current mapping methods, our method of incorporating growth trajectories tends to be more powerful and more precise in QTL detection and effect estimation, as demonstrated in an example using mouse F₂ data.

3.2 Model

Instead of assuming that one QTL is responsible for one trait, it is often more realistic to assume the presence of multiple contributing QTLs, and these QTLs can be located on a same linkage group or different linkage groups. The Bayesian approach described in section 3 to section 6 can be easily extended to the situations that there are s QTLs exist, where s is a fixed number greater than 1, so that multiple QTLs can be searched simultaneously. In the case of s-QTL-model, the QTL loci  λ  become a vector
\((\lambda_1, \lambda_2, \ldots, \lambda_s)\). The MCMC algorithm, can be constructed in a similar manner as described in section 4. Elements of \(\lambda\) are modified consecutively, say for the \(k\)th locus, a proposed \(\lambda_k^*\) generated from \(\text{Uniform}(\max(\lambda_{k-1}, (\lambda_k - \delta), \min(\lambda_{k+1}, (\lambda_k + \delta)))\) is accepted with probability

\[
\min \left( \prod_{i=1}^n \frac{\pi(Q_i | \lambda_{k}^*, M_i) \cdot q(\lambda_k^*, \lambda_k)}{\pi(Q_i | \lambda_k, M_i) \cdot q(\lambda_k, \lambda_k^*)} \right) \cdot 1.
\]

(3-1)

The genotype of \(k\)th locus of individual \(i\) can be updated conditioning on the updated \(\lambda\), the observed phenotypic value \(y_i\), and the genotypes for other QTL locus of individual \(i\). And specifically, the full conditional that genotype of \(k\)th locus of individual \(i\) being \(j\), \((j = 0, 1, \ldots, 3^s)\) is:

\[
p_{ij,k} = \frac{\pi(Q_{i,k} = j | y, \Omega, \Sigma, \lambda) \cdot \pi(y_i | \Omega, \Sigma, Q_{i,k} = j, Q_i)}{\sum_{q=1}^{3^s} \pi(Q_{i,k} = q | \lambda) \cdot \pi(y_i | \Omega, \Sigma, Q_{i,k} = q, Q_i)}.
\]

(3-2)

Given the updated QTL genotypes, in a same way as described in section 2.5, we can update the \(3^s\) blocks of \(\Omega\) one at a time. Finally, depending on different assumptions on the covariance structure, we can either update the covariance matrix as a whole by Gibbs sampling, or update the covariance matrix \(\Sigma\) through updating the matrix-structural parameters. In this analysis, I made no assumption to the structure of the covariance matrix. Therefore, a Gibbs sampling step was applied to update the covariance matrix \(\Sigma\) according its full conditional marginal posterior distribution.

One option to determine the value \(s\) is by running different models under the assumption that there are 0, 1, 2, \ldots QTLs respectively, and then comparing them with Bayes factors. The Bayes factor for two models is defined as the ratio of marginal probabilities of \(y\) given the two models:

\[
BF = \frac{\pi(y | \text{model}_1)}{\pi(y | \text{model}_2)}; \quad (3-3)
\]
and assuming the prior for the models is simply uniform, then according to the Bayes rule, this can be further expressed as:

\[ \frac{\pi(model_1 \mid y)}{\pi(model_2 \mid y)} \frac{\pi(model_2)}{\pi(model_1)} = \frac{\pi(model_1 \mid y)}{\pi(model_2 \mid y)}. \] (3-4)

In practice, a Bayes factor larger than 100 can often be regarded as an evidence supporting model 1.

When comparing two or more models, we perform significance tests to reject or accept a certain hypothesis, by a frequentist approach. Alternatively, Bayes factors not only offer us evidence in favor of a certain hypothesis, but also enable us to incorporate the prior information with regard to the hypothesis of interest. For example, we can impose higher dense on the model assuming two QTLs affect the trait, if previous studies suggested so. Unlike frequentist approaches, this external information will affect the result of model selection. Meanwhile, Kass and Raftery [47] provided an approximation between LOD score and the Bayes factor:

\[ LOD \approx -\log_{10}(BF) - \frac{1}{2}(\text{dim}_{M_2} - \text{dim}_{M_1}) \cdot \log_{10}(n). \] (3-5)

Where \( \text{dim}_{M_1} \) and \( \text{dim}_{M_2} \) are the numbers of parameters in model 1 and model 2 respectively; \( n \) is the sample size. Given this approximation, we can compare these two criterions straightforwardly in both directions.

In our problem, it is impossible to evaluate the likelihoods of parameters analytically, and we have to resort to various numerical approximations. One possible recipe is proposed by Newton and Raftery [62], they suggested an empirical estimator of BF given the MCMC samples:

\[ \hat{\pi}(y \mid \text{model}_s) = \frac{N}{\sum_{i=1}^{N} 1/(y \mid Q^{(t)}, \Omega^{(t)}, \Sigma^{(t)})} \] (3-6)

This is an consistent estimator of the Bayes factor, and converges almost surely to the correct value. However, it is generally too unstable to satisfy a central limit Theorem,
especially when the dimensionality of the parameter space is high. In practice, despite the instability of estimator 3–6, it is very easy to calculate. Moreover, by transforming to the LOD score by 3–5, it usually yields results that are accurate enough for the purpose of interpretation. Hence, in this work, I use 3–6 to estimate the Bayes factors and interpret them on the \textit{log} scale 3–5.

3.3 A Worked Example

3.3.1 Material

Vaughn et al. [90] constructed a linkage map with 96 microsatellite markers for 1043 F$_2$ mice (503 males and 540 females) derived from two strains, the Large (LG/J) and Small (SM/J). This map has a total map distance of $\sim 1780$ cM (in Haldane’s units) and an average interval length of $\sim 23$ cM. For each of the 19 linkage groups, around 10% of the marker genotypes were randomly missing. The F$_2$ progeny was measured for their body mass at 10 weekly intervals starting at age 7 days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth, parity and sex [90].

3.3.2 Results

Zhao et al [113] and Zhao et al [114] first analyzed this data set by using functional mapping with maximum-likelihood based methods. Results from his analysis showed that, instead of a single QTL, multiple QTLs may exist affecting the growth pattern of body mass of mice. And the possible location for these QTLs are on chromosome 6, 7 and 10 respectively. In our analysis, we fitted models under different assumptions with regard to total number of QTLs ($M_0, M_1, \ldots, M_s$) separately. Then, the Bayes factors on a logarithmic scale 3–5 are used to compare these models, and hence, to determine the number of QTL.

The Bayesian formulation of the problem requires specifying prior distributions on the set of model parameter $\Omega$, QTL locus $\lambda$ and the covariance matrix $\Sigma$. According to several related studies, some information are available for $\Omega$, and the priors for $\Omega_j$, ($j = 1, 2, \ldots, 3^s$) are given by a multivariate normal, centered at $(30, 10, 0.6)^T$ and with a large
dispersion $diag(9, 4, 1)$. The prior of the QTL location is simply chosen to be a uniform along the entire linkage group, i.e. $[0, D_m]$ for $\lambda_s$. As to the prior for the covariance matrix, I just followed the conventional choice and set it to be inverse-Wishart($R^{-1}, \rho$), where $R$ is given by the sample covariance matrix.

There are two possible assumption for the two-QTL model, 1) these two QTL are located on a same chromosome; or 2) these two QTL are located on different chromosomes. We examined the data under both assumptions for chromosome 6, 7, and 10. As to the MCMC implementation, the length of a Markov chain consisted of 80,000 cycles. After the first 20,000 cycles (burn-in period), the chain was trimmed by keeping sub-samples in every 60 cycles. Thus the posterior samples contained 1000 observations for post-MCMC analysis.

By fitting the single-QTL model, we observed in figure 3-1 that the marginal posterior density of QTL location has one mode by assuming the QTL is located on chromosome 6 or chromosome 10. On the other hand, by assuming the QTL is located on chromosome 7, the marginal posterior density of QTL location is bimodal, but the two modes are very closed to each other. This suggests that, if there are two QTL, it is more likely that they are located on different chromosomes.

Two QTL on a same chromosome The starting values for the two loci were chose to be at roughly $\frac{1}{3}$ and $\frac{1}{2}$ of the total length of a linkage group. Unsurprisingly, the estimated loci are very closed to each other. Besides that, both additive and dominant effects for the two QTL are very closed, which further support our suspicion that at most one QTL located on each chromosome. Statistically, according to the Bayesian model selection criterion, Bayes factor on the logarithmic scale of $M_1 vs. M_2$ was estimated to be $-15.76$, which is substantially support $M_1$.

Two QTL on a different chromosome $M_2'$, is fitted based on the following assumption: there are exactly two QTL located on two different chromosomes controlling the shape of growth curves of mice body weight. And there are 3 possible sub-models
corresponding to 3 possibilities. These are $M_2' - 1$, the 2 QTL are located on chromosome 6 and chromosome 7; or $M_2' - 2$, the 2 QTL are located on chromosome 6 and chromosome 10; or $M_2' - 3$, the 2 QTL are located on chromosome 7 and chromosome 10. In each case, the starting values for the loci were chose to be $\lambda_s = 40cM (s = 1, 2)$ for both chromosomes.

First, let’s consider model $M_2' - 1$. The first locus $\lambda_1$ was estimated at 53.4$cM$ between marker 2 and marker 3 on chromosome 6, and the second locus $\lambda_1$ was estimated at 43.4$cM$ between marker 2 and marker 3 on chromosome 7. The estimated curve parameters for 9 genotypes and their 95% confidence intervals are presented in table 3-1. Based on the these estimated parameters, growth curves for 9 genotypes can be fitted as shown in figure 3-3. Examining the plot, we can see that 9 growth curves are well apart
Table 3-1. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6 and 7. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQWW</th>
<th>QQWw</th>
<th>QQww</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>α</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QQWW</td>
<td>37.03 (36.13, 38.07)</td>
<td>35.56 (34.88, 36.13)</td>
<td>33.51 (32.67, 34.32)</td>
</tr>
<tr>
<td>QQWw</td>
<td>11.61 (11.10, 12.05)</td>
<td>11.07 (10.61, 11.53)</td>
<td>11.22 (10.60, 11.81)</td>
</tr>
<tr>
<td>QQww</td>
<td>0.64 (0.63, 0.66)</td>
<td>0.66 (0.65, 0.67)</td>
<td>0.66 (0.64, 0.67)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QqWW</th>
<th>QqWw</th>
<th>Qqww</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>α</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QqWW</td>
<td>35.54 (34.95, 36.30)</td>
<td>35.42 (34.94, 36.10)</td>
<td>32.92 (32.15, 33.42)</td>
</tr>
<tr>
<td>QqWw</td>
<td>11.47 (11.07, 11.84)</td>
<td>11.18 (10.79, 11.42)</td>
<td>11.36 (11.03, 11.76)</td>
</tr>
<tr>
<td>Qqww</td>
<td>0.66 (0.65, 0.67)</td>
<td>0.65 (0.64, 0.66)</td>
<td>0.65 (0.64, 0.66)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>qqWW</th>
<th>.qqWw</th>
<th>qqqw</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>α</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qqWW</td>
<td>35.82 (34.81, 36.80)</td>
<td>33.43 (32.77, 34.15)</td>
<td>31.70 (30.90, 32.41)</td>
</tr>
<tr>
<td>qqWw</td>
<td>11.53 (11.05, 12.07)</td>
<td>10.67 (10.25, 11.06)</td>
<td>10.18 (9.75, 10.69)</td>
</tr>
<tr>
<td>qqww</td>
<td>0.64 (0.63, 0.65)</td>
<td>0.65 (0.63, 0.66)</td>
<td>0.65 (0.64, 0.67)</td>
</tr>
</tbody>
</table>

**QTL locus1** 73.19 (66.69, 79.84)

**QTL locus2** 43.16 (39.56, 47.48)

from each other, which suggested a two-QTL model might be appropriate. Also, the plot of marginal posteriors of the two loci is illustrated in figure 3-2. Somewhat consistent with their relative effects, the QTL located on chromosome 7 has a higher peek, and hence is heavier concentrated. On the other hand, the marginal posterior distribution for the QTL located on chromosome 6 has a much heavier tail.

Moreover, when 2 QTL are presented, in a similar manner as described before, we can partition time-dependent genotypic value, \( \mu_{jw}(t)(j = 0, 1, 2; w = 0, 1, 2) \), into the over all mean (\( \mu \)), the additive and dominant effects due to two QTL (\( a_1, a_2, d_1, d_2 \)), as well as the interaction between the two QTL (\( i_{aa}, i_{ad}, i_{da}, i_{dd} \)):

\[
\hat{\mu}(t) = \frac{1}{9} (\mu_{22} + \mu_{21} + \mu_{20} + \mu_{12} + \mu_{11} + \mu_{10} + \mu_{02} + \mu_{01} + \mu_{00}), 
\] (3–7)

\[
\hat{a}_1(t) = \frac{1}{2} (\mu_{22} - \mu_{02}), 
\] (3–8)

\[
\hat{a}_2(t) = \frac{1}{2} (\mu_{02} - \mu_{00}), 
\] (3–9)

\[
\hat{d}_1(t) = \frac{1}{2} (\mu_{12} + \mu_{10} - 2\hat{\mu}(t)), 
\] (3–10)
\[ \hat{d}_2(t) = \frac{1}{2}(\mu_{21} + \mu_{01} - 2\hat{\mu}(t)), \]  
(3–11)

\[ \hat{i}_{aa}(t) = \frac{1}{2}(\mu_{22} + \mu_{00} - 2\hat{\mu}(t)), \]  
(3–12)

\[ \hat{i}_{ad}(t) = \frac{1}{2}(\mu_{21} - \mu_{01} - 2\hat{a}_1(t)), \]  
(3–13)

\[ \hat{i}_{da}(t) = \frac{1}{2}(\mu_{12} - \mu_{10} - 2\hat{a}_2(t)), \]  
(3–14)

\[ \hat{i}_{dd}(t) = \mu_{11} - \hat{d}_1(t) - \hat{d}_2(t) - \hat{\mu}(t), \]  
(3–15)

Dynamic changes in additive, dominant, and interaction effects due to the two QTL are displayed in figure 3-4. We observed that the QTL located on chromosome 7 has a much larger additive during the whole growth progress. The dominant effect of the QTL on chromosome 6 is very small, which is almost negligible, while the dominant effect of QTL on chromosome 7 is negative in a small magnitude. One great advantage by using our approach is that we can enumerate multiple QTL simultaneously, and evaluate all types of interaction effects between these QTL. As illustrated in figure , , and , the interaction effects do exists. Among them, the dominant-dominant interaction effect is particularly strong and it even overpowers the additive effect of the QTL located on chromosome 6. The dominant-additive interaction is also strong and the direction changes over time.

Compared to the results produced by fitting a single-QTL model, the marginal posterior distributions of the two loci are considerably more concentrated, hence result in much narrower confidence intervals for the QTL locations. Besides that, the estimated posteriors also have better shapes, they are much smoother, and more importantly, with only a single mode. We also noticed that confidence intervals for curve parameters are slightly narrower than the confidence intervals produced by fitting a single-QTL model. The Bayes factor of M1-1 (a single QTL is located on chromosome 6) vs. M2-,1, (two QTL are located on chromosome 6 and chromosome 7 respectively) was estimated to be 3.48, and the Bayes factor of M1-2 (a single QTL is located on chromosome 7) vs. M2’-1, was estimated be 2.14. Thereby, we would substantially favor M2’ -1 over a single-QTL model.
Also, we explored other possibilities for a two-QTL model by fitting M2'-2 and M2'-3. Figure 3-2 illustrated the estimated marginal posteriors of the two loci, which were assumed to be located on chromosome 6 and 10 or on chromosome 7 and 10. Generally, the estimated QTL locations are consistent with those yielded from fitting the single-QTL model. Also, the length of the confidence intervals remained about the same, which suggested that the precision of the estimates was not significantly improved. Results of estimated curve parameters for all genotypes are displayed in table 3-2 and 3-3. For these
Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 6 and 7.

Fitted growth curves for 9 genotypes are showed in figure 3-5 and 3-6 in both cases. We observed that the estimation precision even dropped a little bit if M2'-2 or M2'-3 was used. Furthermore, logarithmic scaled Bayes factors were estimated and displayed in table 3-4, and these statistical evidence didn’t support these 2 models either. In both cases, we were in favor of a single-QTL model over model M2'-2 or M2'-3.

3.4 Monte Carlo Simulation

In this section, several simulated numerical examples are presented to illustrate the great advantage by using a Bayesian typed functional mapping method. In chapter, we would like to investigate the statistical properties of multi-QTL model. As in chapter 2, we would like to compare the results produced by using our new approach to those results produced by using the conventional maximum-likelihood based functional mapping.
Figure 3-4. Dynamic changes of the additive, dominant and interaction effects due to the two QTL based on a tow-QTL-model.

Table 3-2. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 6 and 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQWW</th>
<th>QQWw</th>
<th>QQww</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)</td>
<td>37.03 (34.39, 37.72)</td>
<td>36.03 (34.97, 37.25)</td>
<td>34.70 (33.19, 35.94)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>11.84 (11.03, 12.74)</td>
<td>11.59 (10.80, 12.13)</td>
<td>11.39 (10.66, 12.16)</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>0.67 (0.64, 0.69)</td>
<td>0.65 (0.63, 0.66)</td>
<td>0.66 (0.64, 0.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QTL genotype</th>
<th>QQWW</th>
<th>QQWw</th>
<th>QQww</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)</td>
<td>35.11 (33.98, 36.19)</td>
<td>35.11 (33.98, 36.19)</td>
<td>33.64 (32.50, 34.81)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>11.54 (11.03, 12.74)</td>
<td>11.59 (10.80, 12.13)</td>
<td>11.39 (10.66, 12.16)</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>0.66 (0.64, 0.68)</td>
<td>0.65 (0.63, 0.66)</td>
<td>0.64 (0.62, 0.66)</td>
</tr>
</tbody>
</table>

| QTL locus1 | 74.31 (67.18, 81.98) | 75.02 (67.00, 80.64) |
Figure 3-5. Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 6 and 10.

Table 3-3. Bayesian estimates of growth curves for the QTL genotypes detected on mouse chromosome 7 and 10. Numbers in the parentheses are the 95% equal-tail confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QTL genotype</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QQWW</td>
<td>QQWw</td>
<td>QQww</td>
<td>QQWw</td>
</tr>
<tr>
<td>α</td>
<td>35.84 (34.28,37.77)</td>
<td>36.59 (35.53,37.87)</td>
<td>34.15 (32.63,35.73)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.52 (10.58,12.56)</td>
<td>11.92 (11.26,12.60)</td>
<td>11.42 (10.57,12.346)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.66 (0.64, 0.69)</td>
<td>0.64 (0.63, 0.66)</td>
<td>0.66 (0.64, 0.69)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>QqWW</td>
<td>QqWw</td>
<td>Qqww</td>
<td>Qqww</td>
</tr>
<tr>
<td>α</td>
<td>36.13 (34.71,37.43)</td>
<td>34.77 (33.69,35.45)</td>
<td>33.62 (32.56,35.57)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.52 (10.79,12.28)</td>
<td>11.11 (10.40,11.91)</td>
<td>10.90 (10.16,11.62)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.65 (0.63, 0.68)</td>
<td>0.65 (0.63, 0.68)</td>
<td>0.66 (0.64, 0.67)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>qqWW</td>
<td>qqWw</td>
<td>qqww</td>
<td>qqww</td>
</tr>
<tr>
<td>α</td>
<td>34.01 (32.56,35.57)</td>
<td>31.62 (30.66,33.05)</td>
<td>32.83 (31.23,34.56)</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>11.38 (10.65, 12.15)</td>
<td>11.26 (10.73, 11.97)</td>
<td>10.57 (9.80, 11.56)</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td>0.66 (0.63, 0.68)</td>
<td>0.67 (0.65, 0.69)</td>
<td>0.64 (0.63, 0.67)</td>
<td></td>
</tr>
<tr>
<td>QTL locus1</td>
<td>43.03 (39.43,46.76)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL locus2</td>
<td>72.53 (64.77,79.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3-6. Fitted growth curves for the 9 QTL genotypes assuming two QTL are located on mouse chromosome 7 and 10.

Table 3-4. Estimated Bayes factors on a logarithmic scale between different models

<table>
<thead>
<tr>
<th>Models being compared</th>
<th>Logarithmic scaled Bayes factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$ vs. $M_{1\text{-}ch6}$</td>
<td>12.91</td>
</tr>
<tr>
<td>$M_0$ vs. $M_{1\text{-}ch7}$</td>
<td>13.47</td>
</tr>
<tr>
<td>$M_0$ vs. $M_{1\text{-}ch10}$</td>
<td>7.99</td>
</tr>
<tr>
<td>$M_{1\text{-}ch6}$ vs. $M_{2\text{-}ch6&amp;7}$</td>
<td>3.48</td>
</tr>
<tr>
<td>$M_{1\text{-}ch7}$ vs. $M_{2\text{-}ch6&amp;7}$</td>
<td>2.14</td>
</tr>
<tr>
<td>$M_{1\text{-}ch6}$ vs. $M_{2\text{-}ch6&amp;10}$</td>
<td>-2.67</td>
</tr>
<tr>
<td>$M_{1\text{-}ch10}$ vs. $M_{2\text{-}ch6&amp;10}$</td>
<td>-2.86</td>
</tr>
<tr>
<td>$M_{1\text{-}ch7}$ vs. $M_{2\text{-}ch7&amp;10}$</td>
<td>-4.31</td>
</tr>
<tr>
<td>$M_{1\text{-}ch10}$ vs. $M_{2\text{-}ch7&amp;10}$</td>
<td>-2.25</td>
</tr>
</tbody>
</table>
I designed these simulation studies according to an actual dog-body-mass experiment. Inspired by the conditions of the real data introduced previously, we carried out simulation studies to investigate the performance and the statistic properties of our proposed multi-QTL model.

**when more than one QTL located on a same linkage group**

In our experience, if two or more QTL are located on a same chromosome, and are closely linked, enumeration of these QTL can be very difficult by using conventional maximum-likelihood based methods. Hence in this simulation study we are particularly interested in investigating the ability of separating closely linked QTL by using our proposed method.

An $F_2$ population with 600 individuals was simulated for a chromosome segment with length 100 cM covered by 11 evenly spaced markers. Two QTL are placed at both 24 cM and 72 cM from the first marker on the left-hand side. For a particular dynamic phenotypic trait, say body mass, ten measurements with equidistant time lag can be simulated for each individual. The simulated data is based on the following assumptions. First, these vectors of measurements are assumed to follow multivariate normal distributions whose mean-vectors are modeled by the logistic growth curve $2^2$. Second, we assume that both QTLs contribute to this particular dynamic trait by affecting the underlying parameters of growth curves, hence 9 distinct sets of curvature parameters $\alpha_j, \beta_j, \gamma_{ij=1}^9$ were preset for each of the 9 QTL genotypes. Finally, we assume that the covariance matrix is common across all individuals with all genotypes.

In the first simulation scenario, true curvature parameters and the true covariance matrix were preset according to the corresponding estimates yielded from the mice data in section 3. In this setting, the heritability ($H^2$) was estimated to be roughly 0.1. Several preliminary runs show that despite for different initial points the constructed Markov chain can reach to its stationary distribution after 10,000 burn-in cycles. After discarding
these initial burn-in periods, MCMC samples are collected for every 60th cycles in the rest of 60,000 cycles, and this gives us a working sample with sample size 1000.

A small scaled replicated simulation showed consistent behavior of our approach. Although 10 replications might not be sufficiently large to allow accurate estimation of statistic power, we can still get a general idea about the performance of our proposed model. For all 10 replications, the true QTL locations were captured by their corresponding 95% confidence intervals; and for 7 times, the Bayes factor comparing $M_1$ and $M_2$ was in favor of $M_2$, i.e. the Two-QTL model. For the purpose of illustration, results from one of those simulations are presented here. The first locus $\lambda_1$ was estimated at 23.92 cM, and the second locus $\lambda_2$ was estimated at 70.12 cM. Results from one simulation for the estimations of all parameters are summarized 3-5. We noticed that even at such a relatively low heritability level, the precision of all estimates from our proposed method was able to maintain at a satisfactory level.
Table 3-5. Results from a simulation study, where $s = 2$ and $H^2 = 0.1$. Numbers in the parentheses are the 95% equal-tail confidence intervals. Numbers in the box bracts are the given values of the parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QTL genotype</th>
<th>QTL genotype</th>
<th>QTL genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QQWW</td>
<td>QQWw</td>
<td>QQww</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>36.63 [37.06] (35.85, 37.51)</td>
<td>35.52 [35.52] (34.44, 36.33)</td>
<td>33.29 [33.50] (32.24, 34.32)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.63 [0.64] (0.62, 0.65)</td>
<td>0.67 [0.66] (0.65, 0.68)</td>
<td>0.66 [0.66] (0.65, 0.67)</td>
</tr>
<tr>
<td></td>
<td>QqWW</td>
<td>QqWw</td>
<td>Qqww</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>35.66 [35.57] (34.53, 36.61)</td>
<td>35.53 [35.46] (34.92, 36.16)</td>
<td>32.15 [32.84] (31.09, 33.04)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.66 [0.66] (0.64, 0.67)</td>
<td>0.64 [0.65] (0.65, 0.65)</td>
<td>0.66 [0.65] (0.65, 0.68)</td>
</tr>
<tr>
<td></td>
<td>qqWW</td>
<td>qqWw</td>
<td>qqww</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>36.03 [35.78] (34.35, 37.80)</td>
<td>33.48 [33.45] (32.47, 34.56)</td>
<td>32.08 [31.71] (31.49, 33.06)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.66 [0.64] (0.64, 0.67)</td>
<td>0.64 [0.65] (0.62, 0.65)</td>
<td>0.66 [0.65] (0.65, 0.68)</td>
</tr>
<tr>
<td>$QTL locus 1$</td>
<td>23.92 [24.00] (20.90, 26.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$QTL locus 2$</td>
<td>70.12 [72.00] (68.55, 72.16)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As a comparison, I carried out another replicated simulation experiment in a same setting. For these replications, QTLs were enumerated by using conventional functional mapping approach. Profiles of LR-test statistics along the linkage group are summarized in figure 3-8, for all 25 simulated independent samples. For the QTL with larger genetic effects, about 90% of these profiles had a large peak within 8cM around the given location. On the other hand, for the QTL with smaller genetic effects, none of these profiles showed peak in the neighborhood of the given location.

In another scenario, I’d like to evaluate the performance of our model when the heritability is higher. First, I preset the true covariance matrix as the corresponding estimates yielded from the mice data in section 3, then curvature parameters were specified so that the heritability was estimated to be 0.4. Again, since results from several independent replicated simulations are broadly consistent with each other, results from only one of the simulation were presented. Posterior densities of the two QTL location are given in figure 3-9, and the resulting estimations of all parameters are summarized Table 3-6. In general, accuracy of all estimates were improved when I increased the heritability. Among all parameters, locations of two QTLs had the greatest improvement, the estimated 95% confidence intervals were as wide as only $\frac{1}{4}$ of the estimated 95% confidence intervals when heritability was 0.1. Finally, we compared fitted growth curves and the given curves for each of the nine genotypes. As displayed in 3-10, our model was able to provide a perfect fit to the growth curves for all genotypes.

In summary, our proposed method provided more accurate estimates to all parameters of interest, and a much greater power to separate closely linked QTL. Most importantly, our model not only enable us to enumerate multiple QTL along the chromosome, but also a way to evaluate the genetic effects of each individual QTL as well as the interaction between different QTL.
Figure 3-7. Estimated marginal posterior distributions of the two loci on a same chromosome in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.1.
Figure 3-8. The profile of the LR between the full and reduced (no QTL) model estimated from the SAD(1) model for body mass growth trajectories in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.1.
Figure 3-9. Estimated marginal posterior distributions of the two loci on a same chromosome in a simulation study. As to the settings of this simulation, we assume that two QTL control body mass growth trajectories of mice, and the heritability is 0.4.
Figure 3-10. Fitted and given growth curves for the 9 QTL genotypes in a simulation study, assuming the heritability is 0.4.
Table 3-6. Results from a simulation study, where $s = 2$ and $H^2 = 0.4$. Numbers in the parentheses are the 95% equal-tail confidence intervals. Numbers in the box bracts are the given values of the parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>QQWW</th>
<th>QQWw</th>
<th>QQww</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>37.38 [37.00]</td>
<td>35.50 [35.50]</td>
<td>35.49 [35.00]</td>
</tr>
<tr>
<td></td>
<td>(36.73, 38.09)</td>
<td>(34.70, 36.33)</td>
<td>(34.67, 36.32)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>9.86 [10.00]</td>
<td>10.54 [10.50]</td>
<td>7.34 [7.50]</td>
</tr>
<tr>
<td></td>
<td>(9.60, 10.21)</td>
<td>(10.04, 10.97)</td>
<td>(6.80, 7.87)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.59 [0.65]</td>
<td>0.64 [0.65]</td>
<td>0.63 [0.63]</td>
</tr>
<tr>
<td></td>
<td>(0.58, 0.60)</td>
<td>(0.63, 0.66)</td>
<td>(0.61, 0.64)</td>
</tr>
</tbody>
</table>

| QTL locus1 | 23.77 [24.00] | 22.62 [25.04] |
| QTL locus2 | 71.12 [72.00] | 71.98 [73.24] |
3.5 Discussion

Given a recent upsurge of interest in the genetic study of biological processes, there is a pressing need for the development of statistical models for unraveling the genetic etiology of quantitative variation in a dynamic trait. QTL mapping, by associating real biological phenotypes with genetic polymorphisms, can provide an unbiased view of the network of gene actions and interactions that build a complex phenotype like growth trajectories [1] [25] [59] [100]. Through an integrated approach, studies can move to characterize the detailed and precise picture of the genetic architecture of any biological process. By incorporating genetic tests into agricultural and biomedical research programs, breeders might have more desirable opportunities to select superior genotypes in practice, and physicians might improve outcomes for patients.

In this chapter, I incorporated Bayesian approaches to mapping interactive or epistatic QTL that govern dynamic processes of a quantitative trait. This model has been examined through simulation studies. Its application to a real example indicates that this model will be practically useful. It can detect significant genetic variants that control biological processes and provide reasonably precise estimates of the genetic parameters.

There are several ways in which our model may be extended. For simplicity, our presentation is based on the interaction between two QTLs. It is likely that the two-QTL interaction model is too simple to characterize genetic variants for quantitative variation. With the foundation for the two-QTL interaction model, this model can be extended to include the high-order interactions among multiple QTLs which are associated with the phenotypic variation. A high-order QTL epistatic model will encounter the problem of many parameters to be estimated, but it can be expected to have potential implications for understanding the origin and evolution of development and the contributions of epistatic effects to evolutionary changes in the process of development.
CHAPTER 4
PROSPECTS

4.1 Introduction

The motivation of this dissertation is to develop a statistical model and algorithm for mapping genes or quantitative trait loci (QTLs) that regulate dynamic or longitudinal traits, aimed to draw a detailed picture of the genetic architecture of complex biological processes. The new model was derived within the Bayesian context, implemented with Markov chain Monte Carlo (MCMC) algorithms, and integrated with the principle of functional mapping to approach many high-dimensional and multivariate issues characterized by systems biology. In this dissertation, I have, for the first time, constructed a general Bayesian framework for mapping complex dynamic traits, thus increasing the application and visibility of functional QTL mapping in genomic research.

In the last chapter of this dissertation, I will pinpoint several key areas in which Bayesian approaches can be expanded to better understand the genetics of dynamic traits. Some of these areas are biological, some are statistical and many others are mixed.

4.2 Parametric, Nonparametric and Semiparametric Modeling

The model developed in this dissertation is based on growth equation. With the incorporation of other mathematical functions, the new model can find its immediate applications in different areas. The following mathematical functions exist to describe different biomedical processes:

a. Sigmoid curve for tumor growth,
b. Bi-exponential equations for HIV dynamics,
c. Sigmoid Emax models for pharmacodynamic response,
d. Fourier series approximations for periodic cell cycles,
e. Biological thermal dynamics,
d. Ordinal differential equations for biological clock,
e. Fourier series approximation for periodic cell cycle.
By estimating the mathematical parameters that define these different curves, functional mapping allows for the test of genetic control over developmental patterns of variation for a dynamic trait. Likewise, models for the covariance structure propose that a functional relationship exists between the variance or covariance of any two observations and the times of their measurements and possibly other covariates. The parametric covariance functions have been discussed and reviewed by and colleagues [48] [71]. Statistical tests are performed to make an inference about the significance of models for the mean and covariance structures.

Key steps for constructing a general framework for the genetic analysis of biological and biomedical processes have been achieved mostly through parametric approaches. In fact, we can also construct such a framework within nonparametric and semiparametric contexts to model both the mean and covariance structures.

When no functional relationship exists, nonparametric analogues of parametric modeling can be used. Nonparametric regression methods using kernel estimators have been considered for the mean structure of growth curve data. All of these nonparametric approaches have in common that the unknown mean response curve over time is estimated by smoothing the raw data, and time is the only explanatory variable. [61] applied nonparametric regression methods to longitudinal data but without considering a serial correlation structure.

An approach based on B-spline basis functions can be used for nonparametric regression fitting. The B-spline approach constructs curves from pieces of lower degree polynomials smoothed at selected pointed (knots). Brown et al. [7] extended the B-spline basis to model multiple longitudinal variables. The idea of B-spline curve fitting will be incorporated into the functional mapping model, aimed to increase the breadth of the use of functional mapping in solving practical genetic problems. Other applicable dimension reduction methods include functional principal components analysis that is data-driven and lets data speak for themselves.
Relative to nonparametric modeling of the mean structure, nonparametric covariance modeling has received little attention. Diggle and Verbyla [18] used kernel-weighted local linear regression smoothing of sample variograms ordinates and of squared residuals to provide a nonparametric estimator for the covariance structure without assuming stationarity. In addition, they used the value of the estimator as a diagnostic tool but did not study the use of the estimator in more formal statistical inference concerning the mean profiles. Wang [97] used kernel estimators to estimate covariance functions in a nonparametric way. His only assumption was to have a fully unstructured smooth covariance structure, together with a fixed effects model. The proposed kernel estimator was consistent with complete but irregularly spaced follow-ups, or when the missing mechanism is strongly ignorable MAR (missing at random).

Zeger and Diggle [116] and Moyeed and Diggle [60] studied a semiparametric model for longitudinal data in which the covariates entered parametrically and only the time effect entered nonparametrically. To fit the model, they extended to longitudinal data the backfitting algorithm of Hastie and Tibshirani for semiparametric regression. Based on the idea used in semiparametric mean response models, we can extend the semiparametric models for functional mapping to allow for more flexibilities in both mean and covariance structures.

4.3 Toward a Comprehensive Biology

Functional mapping is flexible for any extension to consider allometric scaling of multiple biological variables, irregular shapes of abnormal organ, multiple environmental variables, high-dimensional responses and ontogenetic reaction norms that are all essential to construct a comprehensive picture of biology. The following include specific issues that should be addressed and can be addressed within the Bayesian paradigm.

4.3.1 Joint Models of Longitudinal Trajectories and Time-to-Events

In pharmacogenetics, identification of specific genetic variants responsible for longitudinal trajectories (such as HIV dynamics) and time-to-events (such as the time
to onset of AIDS symptoms) can help to design individualized drugs to control patient’s progression to a disease. Similarly, a shared genetic basis between prostate specific antigen, repeatedly measured for patients following treatment for prostate cancer, and the time to disease recurrence can be used to make optimal treatment schedules. Reproductive plant behaviors, such as the time to first flower and the time to form seeds, may be associated with growth rates and sizes of plants, which is regarded as the consequence of plant’s adaptation to the environment in which they are grown. These so-called time-to-events can be incorporated into Bayesian-based functional mapping, with the assumption that they are controlled by QTLs that regulate developmental processes.

4.3.2 Multivariate Longitudinal Trajectories

Understanding the genetic architecture of different biological traits and their co-regulation mechanisms during ontogeny is fundamental to quantitative developmental genetics. Traditionally, this knowledge has been thought to be crucial for the study of evolutionary biology and design of plant and animal breeding. With a continuously increasing demand in these fields, the model proposed in this dissertation should be extended to include multiple biological traits, ultimately providing a useful tool to shed light on the genetic basis for various integrated developmental features of biology.

4.4 Statistical Considerations

4.4.1 Model Selection

Previously, I describe the details of model selection by fixing s, the number of QTL. Although this model selection procedure is technically straightforward, it has certain downside. First, it generally requires high computational demands to fit different models. Also, Bayes factors, the criterion of model selection I used can be sometimes unstable. Finally, I did not incorporate any possible prior information about the number of QTL into the process of model selection.

A feasible alternative way to determine the number of QTL would be an application of the reversible-jump algorithm. Theoretically, if we regard the number of QTLs, i.e.,
s itself as an unknown parameter, and an reversible jump MCMC algorithm [33] can be utilized to determine the value of s. For a ‘death move’ in a reversible jump MCMC, a selected QTL loci is removed with a particular probability ($\alpha_d$) and the value of s is deducted to $s - 1$; on the other hand, s may also be increased to $s + 1$ in a ‘birth move’. In a birth move, a newly generated candidate QTL loci will be included in the model for a certain probability ($\alpha_b$), and a new set of curvature parameters $\Omega$ need to be proposed.

However, a direct implementation of reversible jump MCMC algorithm may encounter great difficulties. This is due to 1) the dimension of parameter space change greatly when s changes, even when s is as small as 2) computing time is usually lengthy because of slow-mixing 3) convergence of the algorithm strongly depended on a good propose. Hence, there is a lot of room left for improvement. First, Descending-dimension methods can be explored to decrease the dimension of parameter space especially when s is large. Second, the acceptance probability of a birth step can be adjusted to prevent too much jumping between models.

### 4.4.2 Sensitivity Studies and Other Estimation Issues

In theory, the Metropolis-Hastings algorithm guarantees the convergence of samples from the Markov chain to the corresponding target distribution under moderate regularity conditions. But many question remain in practice: (1) How long should this chain take to reach its stationary distribution? In other words, what is an appropriate burn-in for this type of problem? (2) How long should the chain proceed to assure the sufficient accuracy of the produced estimators, especially when the dimension of parameter space is very large? These questions can be addressed through extensive simulation studies, aimed to further explore the stability and the power of the Bayesian approach.

### 4.4.3 Testing QTL Effects

Although the estimated posteriors of QTL locations have notable bumps, their QTL effects may not be significant. And even for a detected QTL, there is none statistical test procedure available to evaluate the significance of its genetic effects. For this reason,
One of the criticism of Bayesian QTL mapping methods might be the lack of statistical testing on the detected QTL. In the case of functional mapping, the evaluation of QTL effects can be very challenging. This is because the expected curve of a longitudinal trait is determined by several curvature parameters simultaneously. Therefore, significant difference of individual curvature parameters does not necessarily result in significance difference of curves, and vice versa. Therefore, it would be very important and meaningful to develop a statistical testing procedure to assess the effects of the detected QTLs.

4.5 Conclusions

With considerable achievements in genetic mapping due to collective efforts of researchers in the past two decades [42] [50] [65] [69] [70] [107] [115] it is now possible to detail the genetic architecture of a complex, quantitatively inherited trait. In her seminal review, Mackay [55] defined the overall picture of the genetic architecture of a complex trait in terms of its composed elements, i.e., the number of genes involved and their frequencies and pleiotropic effects, gene-gene (epistatic), gene-sex and gene-environment interactions. In a follow-up review, Mackay and colleague [56] further documented the importance of these elements in creating and maintaining the genetic variation of a specific quantitative trait in a population. Wu and colleagues ([54] and [107]) pioneered a general statistical framework for functional mapping that can be used to unravel the genetic architecture of dynamic complex traits, as defined by Mackay [56].

In this dissertation, I developed a general Bayesian framework for functional mapping of complex dynamic traits. Compared with traditional maximum likelihood-based functional mapping, the Bayesian approach displays several favorable properties in parameter estimation and the characterization of the number of the underlying genes or QTLs for the trait. The understanding of the QTL number involved is a key to draw a comprehensive picture of the genetic network and regulation that determine the developmental shape and process of a complex trait. With the availability of a complete reference sequence of the entire genome for various species and continuing advances in
powerful statistical tools like one derived in this dissertation, we will be in an excellent position to help geneticists, breeders and clinical doctors extract useful genetic information from the data sets that are often high-dimensional and have complex dynamic structure.
REFERENCES


[22] D. S. Falconer and T. F. C. Mackay, Introduction to quantitative genetics, 4th

Hall New York 1996.


Meller, R. Elber, K. B. Alpert and S. D. Tanksley, “A quantitative trait locus key to


[28] A. Gallavotti, Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley,
M. E. Pe and R. J. Schmidt, “The role of barren stalk1 in the architecture of

[29] S. Geman and D. Geman “Stochastic relaxation, Gibbs distributions, and the


BIOGRAPHICAL SKETCH

Tian Liu was born on April 1, 1979, in Shanghai, China. She majored in applied mathematics at Fudan University (Shanghai, China), and obtained her Bachelor of Science in 2001. During 2001 and 2002 she worked as a high school math teacher in Kingsford Community School, London, U.K. From 2002 to present, she has been studying in the Department of Statistics at the University of Florida. She received her Master of Science in statistics in 2004. She received her Ph.D. in 2007. She will work in the Genome Institute of Singapore as a research scientist.