FUNCTIONAL MAPPING OF DYNAMIC SYSTEMS

By

JIANGTAO LUO

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

2009

I dedicate this to everyone who has helped me

ACKNOWLEDGMENTS

I am grateful to my supervisory committee members Dr. Hartmut Derendorf, Dr. Jayadeep Gopalakrishnan, and Dr. Ronald Randles for their valuable critiques, suggestions, and remarks. Special thanks are given to my supervisors Dr. William Hager and Dr. Rongling Wu for their guidance and great help. In the past years they have distilled their research experience to me in every possible way. Every step of my progress has slated their hard work. I am very indebted to Dung Phan and Dr. Xiqiang Zheng for their unselfish help. I am deeply obligated to Professor David Wilson for his help and encouragement. Thanks are given to my wife Yanpin and my son Binjie. I owe much to their efforts, upon which I rely. On the heaven I want to thank my grandmother, mother, and brother since they sacrificed everything they had for my education. Finally, I thank my father for supporting my educations in my early life and providing me stubbornness to finish my graduate school.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 9

LIST OF SYMBOLS 10

ABSTRACT 11

CHAPTER

1 INTRODUCTION 13

1.1 Problems 13
1.2 Genetic Control of Phenotypes 16
1.3 Dissertation Goals 18

2 A DIFFERENTIAL EQUATION MODEL FOR FUNCTIONAL MAPPING OF A VIRUS-CELL DYNAMIC SYSTEMS 19

2.1 Introduction 19
2.2 Dynamic Models of Virus Load 20
2.2.1 Differential Equations 20
2.2.2 Markov Properties 21
2.3 Functional Mapping 23
2.3.1 Genetic Design 23
2.3.2 Likelihood 24
2.3.3 Estimation and Algorithm 26
2.4 Hypothesis Testing 27
2.4.1 The Significance of QTL 27
2.4.2 Genetic Mechanisms 28
2.4.3 Physiological Control of QTL 28
2.5 Application to Simulated Data 29
2.6 Discussion 31

3 BAYESIAN INFERENCE FOR GENETIC MAPPINGS OF DRUG RESISTANCE 37

3.1 Introduction 37
3.2 Dynamic Models of Drug Resistance 38
3.2.1 Differential Equations 38
3.2.2 Bayesian Markov Model for Drug Resistance 39
3.3 Genetic Mapping for Drug Resistance 42
3.3.1 Likelihood Function 42

3.3.2 Estimation and Algorithm . . . 44
3.3.3 Testing the Significance of QTL and Drug Resistance . . . 45
3.4 Computer Simulation and Discussion . . . 45

4 EM ALGORITHM FOR SOLVING MIXTURE MODELS IN COMPLEX GENETIC TRAITS . . . 50

4.1 Introduction . . . 50
4.2 Algorithm . . . 50
4.2.1 Genetic Design . . . 50
4.2.2 Likelihood . . . 51
4.2.3 Algorithm . . . 53
4.2.4 Main Results . . . 61
4.3 DSE Data Analysis . . . 71
4.4 Numerical Experiment . . . 72
4.5 Conclusion and Discussion . . . 74

5 ONGOING RESEARCH AND FUTURE WORK . . . 75

REFERENCES . . . 77

BIOGRAPHICAL SKETCH . . . 87

LIST OF TABLES

Table page

2-1 Joint genotype frequencies at the marker and QTL in terms of gametic haplotype frequencies, from which the conditional probabilities of QTL genotypes given marker genotypes can be calculated according to Bayes' theorem. 33

2-2 The MLEs of parameters that define virus-host dynamics for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 34

2-3 The MLEs of parameters that define virus-host dynamics for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.4$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 35

3-1 The MLEs of parameters that define the dynamics of viral drug resistance for three different QTL genotypes and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.05$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 48

3-2 The MLEs of parameters that define the dynamics of viral drug resistance for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 49

4-1 Estimates of the Parameters for five codons in DSE data 72

4-2 The MLEs of parameters for three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 73

4-3 The MLEs of parameters for three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.4$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs. 73

4-4 Comparison the MLEs of Newton method and EM algorithm for Estimating the parameters of three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.1$. The number in the parentheses are the computation errors. 74

LIST OF FIGURES

Figure page

1-1 Leaf, stem, and root parts of a plant. M: biomass, L: length, D: cross-sectional area, ρ: porosity of roots and stem. Adapted from [132] by Zens and Webb (2002). 15

2-1 Estimated and true curves for a system of viral infection including uninfected cells 36

LIST OF SYMBOLS, NOMENCLATURE, OR ABBREVIATIONS

$\sum$	Denotes the summation of a series of terms
$\prod$	Denotes the product of a series of terms
Gene	DNA segment needed to make a protein.
Allele	Different version of the same gene.
Genotype	All alleles of an individual.
Genetic Mapping	Any method for discovering the location of genes related to a phenotype using recombination.
Linkage or Genetic Linkage	The behavior of the genes on the same chromosome often do not assort independently.
Linkage Disequilibrium	The non-random association of alleles at two or more loci.
Locus	The site on a chromosome where a gene is located.
Phenotype	All traits of an organism (human), especially the traits we study.
Marker or Genetic Marker	A gene or DNA segment with known location for a particular trait.

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

FUNCTIONAL MAPPING OF DYNAMIC SYSTEMS

By

Jiangtao Luo

December 2009

Chair: William Hager
Cochair: Ronling Wu
Major: Mathematics and Statistics

The dynamic pattern of viral load in a patient's body critically depends on the host's genes. For this reason, the identification of those genes responsible for virus dynamics, although difficult, is of fundamental importance to design an optimal drug therapy based on patients' genetic makeup. Here, we present a differential equation (DE) model for characterizing specific genes or quantitative trait loci (QTLs) that affect viral load trajectories within the framework of a dynamic system. The model is formulated with the principle of functional mapping, originally derived to map dynamic QTLs, and implemented with a Markov chain process. The DE-integrated model enhances the mathematical robustness of functional mapping, its quantitative prediction about the temporal pattern of genetic expression, and therefore its practical utilization and effectiveness for gene discovery in clinical settings. The model was used to analyze simulated data for viral dynamics, aimed to investigate its statistical properties and validate its usefulness. With an increasing availability of genetic polymorphic data, the model will have great implications for probing the molecular genetic mechanism of virus dynamics and disease progression. This thesis consists of five chapters. In Chapter 1 we briefly summarize the importance of study the dynamic system from the genetic viewpoint. In Chapter 2 we develop the general framework for virus dynamic models. We focus on drug resistance with parameters having Bayesian structure in Chapter 3. Chapter 4 discusses the EM algorithm of mixture models used in Chapters 2 and

3. Some useful results have been given with strict math proof, which guarantees the correctness of the algorithm. The final chapter, Chapter 5, we talk about the ongoing research and future work.

CHAPTER 1
INTRODUCTION

1.1 Problems

Many traits of agricultural, biological and biomedical importance are difficult to study because they are often under the control of many genes each with a small effect in an interactive manner. This situation has now been changed with the emergence of sequencing and genotyping technologies that allow the generation of DNA-based marker data, crucial for the genetic mapping of complex traits, almost with no limit. These technical advances appeal to the development of powerful statistical methods that can analyze the mapping data effectively and efficiently. The past two decades have witnessed a dramatic growth of statistical methodologies suited for various types of mapping populations and markers (see [9, 41, 43, 49, 67, 78, 97, 125, 129, 131, 139]). A new challenge that now faces methodological development is how to dissect a phenotypic trait into its biological components and then reorganize these components into a new phenotype beneficial to humans.

(a) **From static to dynamic mapping**: One of the most interesting topics in genetic studies is to use and develop dynamic models to compare the differences of genetic control at different stages of complex traits (see [42, 45, 46, 64, 71, 72, 82, 126]). Unlike the traditional static models that analyze phenotypic traits at individual time points, the central motivation of dynamic models lies in the study of the temporal pattern of genetic variation for a quantitative trait in a time course [1] and the identification of specific genes (i.e., quantitative trait loci or QTLs) that determine such a time-dependent change of the trait [113]. These models have been instrumental for detecting and mapping dynamic QTLs for individuals traits, such as stem growth and root growth in forest trees [120, 133], plant height in rice [137], tiller number increase in rice [20], biomass growth in soybeans [54], body mass growth in mice [117, 136], body height growth in humans [52], and drug response [57].

(b) **Understanding phenotypes as a dynamical system**: The formation of any phenotypic trait undergoes complex interactions and coordination of its different components expressed at various organizational levels from cell to tissue to organ to organism. A full understanding of these interactive relationships among components may help shed light on the components of the biological systems and predict physiological and pathological states of the systems. This has been feasible by developing a system of differential equations that describe the dynamic behavior and coordination of the biological system based on natural laws. Below is shown the two typical examples for system dissection and modeling:

Example 1 Biomass partitioning: Plant biomass growth is not simply the addition of individual parts (leaves, stem, and roots; Figure 1-1); more importantly, it entails the coordination of these parts through natural laws. These laws include maximizing leaf surface area for photosynthesis and minimizing the transport distance for water, nutrients, and carbon. The coordination of leaf, stem, and root biomass for a plant can be described by a system of ordinary differential equations (ODEs):

$$\begin{aligned} \frac{dM_L}{dt} &= \alpha_L M_T^{\beta_L} - \lambda_L M_L, \\ \frac{dM_S}{dt} &= \alpha_S M_T^{\beta_S}, \\ \frac{dM_R}{dt} &= \alpha_R M_T^{R} - \lambda_R M_R. \end{aligned} \tag{1–1}$$

where M_L , M_S, and M_R are the biomasses of the leaves (L), the stems (S), and the roots (R), respectively, with whole-plant biomass $M_T = M_L + M_S + M_R$, α and β are the constant and exponent power of an organ biomass scaling as whole-plant biomass, and λ is the rate of eliminating ageing leaves and roots [12].

The interactions between different parts of a plant can be modeled and studied by estimating and testing the ODE parameters $(\alpha_L, \beta_L, \lambda_L, \alpha_S, \beta_S, \alpha_R, \beta_R, \lambda_R)$.

[htb]

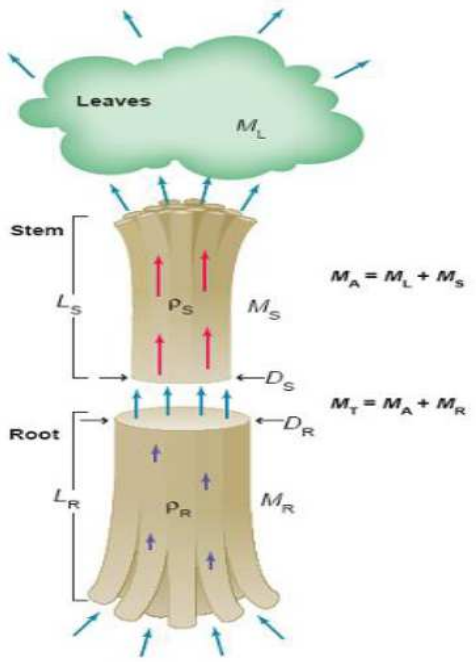

Figure 1-1. Leaf, stem, and root parts of a plant. M: biomass, L: length, D: cross-sectional area, ρ: porosity of roots and stem. Adapted from [132] by Zens and Webb (2002).

Example 2 Cell-viral system: How does viral load change in a patient/s body after an antiviral drug is administrated? This process constitutes a complex dynamical system in which different types of viral cells, including uninfected cells (x), infected cells (y), and free virus particles (v), interact with each other to determine the pattern of viral load in response to drugs ([35] by Ho et al. 1995; [80, 81] by Perelson et al. 1997, 1996; [96] by Sedaghat et al. 2008). A basic model for describing short-term virus dynamics (see [5] by Bonhoeffer et al. 1997) is expressed as

$$\begin{aligned} \frac{dx}{dt} &= \lambda - dx - \beta xv \\ \frac{dy}{dt} &= \beta xv - ay \\ \frac{dv}{dt} &= ky - uv, \end{aligned} \tag{1–2}$$

where uninfected cells are yielded at a constant rate, λ, and die at the rate dx; free viruses infect uninfected cells to yield infected cells with rate βxv; infected cells die with

rate ay; and new virus is yielded from infected cells with rate ky and dies with rate uv. The dynamic pattern of this system can be determined and predicted by the change of these parameters $(\lambda, d, \beta, a, k, u)$.

1.2 Genetic Control of Phenotypes

Existing dynamical system models can map QTLs for genetic variation in the trajectory of a dynamic trait among individuals. If different genotypes at a given QTL correspond to different shapes of curves, this QTL is thought to play a role in governing differentiation in trait trajectories. Thus, by estimating the curve parameters for each QTL genotype and testing the differences in these parameters among genotypes, we can determine whether a dynamic QTL exists and if so, how it triggers an effect on the formation process of a trait. New models based on dynamical systems, such as (1–1) and (1–2), will be able to map QTLs for trajectories of individual traits as well as interactions among different components of the system. Below is a description of how QTL mapping within a dynamical system can be used to address fundamental biological questions:

(a) Size-shape relationship: Size does matter, but shape may matter even more in nature. Shape is one of the most conspicuous aspects of an organism/s phenotype and provides an intricate link between biological structure and function in changing environments. Given the parameters $(\alpha_L, \beta_L, \lambda_L, \alpha_S, \beta_S, \alpha_R, \beta_R, \lambda_R)$ for system (1–1), one can see how much biomass has been allocated to the leaves, stem, and roots. It is possible that some plants have a dominant main stem, with less leaves, while some plants allocate more carbon to the roots (below ground) than the leaves and stem (above-ground). Thus, by integrating the ODE (1–1) into a QTL mapping framework, specific effects of a QTL on a plant/s size and form or shape can be estimated. Furthermore, how the QTL governs the dynamic relationship between size and shape can be quantified.

(b) Structural-functional relationship: There has been a long-standing interest in understanding the relationships between structure and function. The change of structure for a system will quickly lead to the alteration of function. For a plant in drought soil, more energy should be allocated into the root system in order to increase its survival rate and fitness. If the ODE (1–1) is implemented with an additive fitness variable, this will constitute a dynamical system for structural-functional relationships. Genetic mapping of QTLs for such relationships will shed light on the genetic mechanisms involved in balancing vegetative and reproductive growth.

(c) Cause-effect relationship: A web of directed events forms a complex cause-effect relationship. The use of an antiviral drug can increase the amount of uninfected cells by reducing the load of free virus particles in a patient, which reduces the likelihood of the patient to progress into AIDS. Such cause-effect relationships between different types of cells can be quantified by differential equations (1–2). Integrated with QTL mapping models, one can determine how specific QTLs control the dynamic changes of different types of cells in the course of time.

(e) Sink-source relationship: In plants, the function of carbohydrate source to sink relationships determines their productivity. Carbohydrates are transported from supply areas (sources) to areas of growth or storage (sinks). Carbohydrates are produced through photosynthesis in the leaves and channeled through the phloem to the roots, which act as the main carbohydrate sinks during growth. The rate of carbohydrate transport is primarily ruled by the sink strength of plant organs. A dynamic system of sink-sources relationships is composed of potential growth rate, carbon losses through growth and maintenance respiration processes, and carbon demand related to active reserve storage. The identification of specific QTLs that affect these components and therefore sink-sources relationships can be made possible by constructing a system of ODEs and integrating it with the principle of QTL mapping.

1.3 Dissertation Goals

The motivation of this dissertation comes from a pressing need for mathematical and statistical tools that map the genetic architecture of complex phenotypes important to agriculture, biology, and biomedicine. The specific goals of this dissertation include

(1) Incorporate sophisticated differential equations into statistical models for QTL mapping, allowing the understanding of biological problems of high complexity;

(2) Implement rigorous biological principles into the statistical infrastructure, enabling biologically meaningful interpretations of the models and application for real data analysis.

(3) Develop computational algorithms to handle various complexities of genetic and phenotypic data. The EM algorithm coupled with various numerical methods for solving differential equations will be used in estimation and testing, broadening the applications of general statistical theory.

CHAPTER 2
A DIFFERENTIAL EQUATION MODEL FOR FUNCTIONAL MAPPING OF A VIRUS-CELL DYNAMIC SYSTEMS

2.1 Introduction

Several serious human diseases, such as AIDS, hepatitis B, influenza, and rabies, are caused by viruses. To control these diseases, antiviral drugs have been developed to prevent infection of new viral cells or stop already-infected cells from producing infectious virus particles by inhibiting specific viral enzymes. This process constitutes a complex dynamic system, in which different types of viral cells, including uninfected cells, infected cells, and free virus particles, interact with each other to determine the pattern of viral change in response to drugs [1–5]. A major challenge that faces drug development and delivery for controlling viral diseases is to develop a quantitative model for analyzing and predicting the dynamics of decline in virus load during drug therapy and further providing estimates of the rate of emergence of resistant virus.

The development of such a model can now be made possible with recent advances in two seemingly unrelated areas. First, the combination between novel instruments and an increasing understanding of molecular genetics has led to the birth of high-throughput genotyping assays for single nucleotide polymorphisms (SNPs). Through the construction of a haplotype map (HapMap) with SNP data [6], we are able to characterize concrete nucleotides or their combinations that encode a complex phenotype, and ultimately document, map a nd understand the structure and patterns of the human genome linked to drug response. Second, the past two decades have witnessed a tremendous growth of interest in deriving sophisticated mathematical models for characterizing virus dynamics from molecular and cellular mechanisms of interactions between virus and drug [1–5,7–9]. These models mostly built with differential equations have been instrumental for studying the function of virus and the origins and properties of virus dynamics.

These two advances can be integrated to identify specific genes or quantitative trait loci (QTLs) that regulate a dynamic system of viral infection through a new statistical model called functional mapping [10–15]. The basic idea of functional mapping is to map dynamic QTL for the pattern of developmental changes in time course. The purpose of this article is to propose a statistical strategy for implementing a system of differential equations into the functional mapping framework, ultimately to map QTLs from the host genome that determine the dynamic pattern of virus load in patients' bodies. The new strategy is founded on a set of random samples drawn from a natural population at Hardy-Weinberg equilibrium. We integrate the Markov chain properties of dynamic data into the model to facilitate the estimation of parameters that define virus dynamics. Simulation studies were performed to investigate statistical properties of the model and validate its usefulness and utilization.

2.2 Dynamic Models of Virus Load

2.2.1 Differential Equations

A basic model for describing short-term virus dynamics was provided by many researchers [7–9]. This model includes three variables: uninfected cells, x, infected cells, y, and free virus particles, v. These three types of cells interact with each other to determine the dynamic changes of virus in a host's body, which can be described by a system of ordinary differential equations (ODE):

$$\begin{aligned} \frac{dx}{dt} &= \lambda - dx - \beta xv \\ \frac{dy}{dt} &= \beta xv - ay \\ \frac{dv}{dt} &= ky - uv, \end{aligned} \tag{2–1}$$

where uninfected cells are yielded at a constant rate, λ, and die at the rate dx; free virus infects uninfected cells to yield infected cells at rate βxv; infected cells die at rate ay; and new virus is yielded from infected cells at rate ky and dies at rate uv [8]. The system (2–1) is defined by six parameters $\{\lambda, d, \beta, a, k, u\}$ and the initial conditions for

x, y, and v. The dynamic pattern of this system can be determined and predicted by the change of these parameters and the initial conditions of x, y, and v. There are some practical problems in the real application. First, we can only observe the data for x, y, and v at discrete time points, and it is difficult to get the continuous $\frac{dx}{dt}$, $\frac{dy}{dt}$, and $\frac{dv}{dt}$ terms. Second, any biological development is related to genes, but the model does not involve any genetic components. Third, the dynamic change of the virus is accompanied by noise which cannot be neglected in the dynamic modeling. Fourth, this random noise or development noise will be carried from one stage to the next. It should be noted that the model (2–1) used to explain our idea in this article is a basic sculpture of real virus infection as it ignores the dynamics of immune responses and virus mutations.

Let $0 = t_0 < t_1 < \ldots < t_N = T$ denote a mesh on the time interval $[0, T]$ and define $\Delta t_k = t_{k+1} - t_k$. The Euler approximation to the continuous differential equations (2–1) is

$$\begin{aligned} \frac{x(t_{k+1}) - x(t_k)}{\Delta t_k} &= \lambda - dx(t_k) - \beta x(t_k)v(t_k) \\ \frac{y(t_{k+1}) - y(t_k)}{\Delta t_k} &= \beta x(t_k)v(t_k) - ay(t_k) \\ \frac{v(t_{k+1}) - v(t_k)}{\Delta t_k} &= ky(t_k) - uv(t_k), \end{aligned} \tag{2–2}$$

or equivalently,

$$\begin{aligned} x(t_{k+1}) &= x(t_k) + \lambda\Delta t_k - dx(t_k)\Delta t_k - \beta x(t_k)v(t_k)\Delta t_k \\ y(t_{k+1}) &= y(t_k) + \beta x(t_k)v(t_k)\Delta t_k - ay(t_k)\Delta t_k \\ v(t_{k+1}) &= v(t_k) + ky(t_k)\Delta t_k - uv(t_k)\Delta t_k. \end{aligned} \tag{2–3}$$

2.2.2 Markov Properties

Suppose there is a random sample with n patients from a population carrying a certain virus. Each patient is measured for uninfected cells, x, infected cells, y, and free virus particles, v, at a series of time points, $(t_0, t_1, \ldots, t_N)$. Thus, three sets of serial measurements are expressed as $\mathbf{x}_i = [x_i(t_0), \ldots, x_i(t_N)]$, $\mathbf{y}_i = [y_i(t_0), \ldots, y_i(t_N)]$,

and $\mathbf{v}_i = [v_i(t_0), \ldots, v_i(t_N)]$, where the subscript i corresponds to the patient and t_j, $0 \leq j \leq N$, are the measurement times.

A Markov transitional model is used to describe the random process of the system by

$$\begin{aligned} x_i(t_{k+1}) &= x_i(t_k) + \lambda \Delta t_k - d x_i(t_k) \Delta t_k - \beta x_i(t_k) v_i(t_k) \Delta t_k + \epsilon_{x_i}(t_k) \\ y_i(t_{k+1}) &= y_i(t_k) + \beta x_i(t_k) v_i(t_k) \Delta t_k - a y_i(t_k) \Delta t_k + \epsilon_{y_i}(t_k) \\ v_i(t_{k+1}) &= v_i(t_k) + k y_i(t_k) \Delta t_k - u v_i(t_k) \Delta t_k + \epsilon_{v_i}(t_k), \end{aligned} \tag{2–4}$$

where $\epsilon_{x_i}(t_k) \sim N(0, \sigma_x^2)$, $\epsilon_{y_i}(t_k) \sim N(0, \sigma_y^2)$, and $\epsilon_{v_i}(t_k) \sim N(0, \sigma_v^2)$ are the innovation errors for three variables, *x*, *y*, and *v*, respectively, each of which is assumed to be iid and time-independent. To simplify our line of analysis, we assume that these three variables are independent of each other, although this assumption can be relaxed.

For simplicity, we use x_{ik}, y_{ik}, and v_{ik} to stand for $x_i(t_k)$, $y_i(t_k)$, and $v_i(t_k)$, respectively. For a conditional density function, $f(.|.)$, we derive the Markov properties of the dynamic system (2–1) as follows:

Theorem 1.1: *All the future values of uninfected cells, infected cells, and free virus particles depend statistically only on their present values*. That is,

$$\begin{aligned} &f(x_{ik+1}, y_{ik+1}, v_{ik+1} | (x_{i1}, y_{i1}, v_{i1}), \ldots, (x_{ik}, y_{ik}, v_{ik})) = f(x_{ik+1}, y_{ik+1}, v_{ik+1} | (x_{ik}, y_{ik}, v_{ik})), \\ &f(x_{ik+1} | (x_{i1}, y_{i1}, v_{i1}), \ldots, (x_{ik}, y_{ik}, v_{ik})) = f(x_{ik+1} | (x_{ik}, y_{ik}, v_{ik})), \\ &f(y_{ik+1} | (x_{i1}, y_{i1}, v_{i1}), \ldots, (x_{ik}, y_{ik}, v_{ik})) = f(y_{ik+1} | (x_{ik}, y_{ik}, v_{ik})), \\ &f(v_{ik+1} | (x_{i1}, y_{i1}, v_{i1}), \ldots, (x_{ik}, y_{ik}, v_{ik})) = f(v_{ik+1} | (x_{ik}, y_{ik}, v_{ik})). \end{aligned}$$

The proof follows directly from (2–4) and the definitions of $\epsilon_{x_i}(t_k)$, $\epsilon_{y_i}(t_k)$, and $\epsilon_{z_i}(t_k)$. From this theorem, we have the following results.

Corollary 1.2.1: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), $(x_{ik-1}, y_{ik-1}, v_{ik-1})$ *and* $(x_{ik+1}, y_{ik+1}, v_{ik+1})$ *are statistically independent.*

Corollary 1.2.2: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), x_{ik-1} *and* x_{ik+1} *are statistically independent.*

Corollary 1.2.3: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), y_{ik-1} *and* y_{ik+1} *are statistically independent.*

Corollary 1.2.4: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), v_{ik-1} *and* v_{ik+1} *are statistically independent.*

Since

$$((x_{ik+1}, y_{ik+1}, v_{ik+1})|(x_{ik}, y_{ik}, v_{ik}), (x_{ik-1}, y_{ik-1}, v_{ik-1})) = f((x_{ik+1}, y_{ik+1}, v_{ik+1})|(x_{ik}, y_{ik}, v_{ik})),$$

conditional on (x_{ik}, y_{ik}, v_{ik}), $(x_{ik-1}, y_{ik-1}, v_{ik-1})$ and $(x_{ik+1}, y_{ik+1}, v_{ik+1})$ are statistically independent [16]. Hence, **Corollary 1.2.1** holds. The proofs of **Corollaries 1.2.2**, **1.2.3**, and **1.2.4** can be made in a similar way.

Now, we get the following theorems:

Corollary 1.2.5: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), (x_{ij}, y_{ij}, v_{ij}) *for* $j = 0, 1, \ldots, k-1$ *and* $(x_{ik+1}, y_{ik+1}, v_{ik+1})$ *are statistically independent.*

Corollary 1.2.6: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), $\{x_{i1}, \ldots, x_{ik-1}\}$, *and* x_{ik+1} *are statistically independent.*

Corollary 1.2.7: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), $\{y_{i1}, \ldots, y_{ik-1}\}$, *and* y_{ik+1} *are statistically independent.*

Corollary 1.2.8: *Conditional on* (x_{ik}, y_{ik}, v_{ik}), $\{v_{i1}, \ldots, v_{ik-1}\}$, *and* v_{ik+1} *are statistically independent.*

All these corollaries will be used to derive computing algorithms for solving a system of differential equations (2–4) embedded in functional mapping.

2.3 Functional Mapping

2.3.1 Genetic Design

Genetic mapping of QTLs can be based on linkage analysis for a pedigree [17] or linkage disequilibrium analysis for a natural population [11]. In this article, we

assume that the population used to map human QTLs for viral load trajectories is composed of n patients randomly sampled from a natural population at Hardy-Weinberg equilibrium (HWE). A panel of SNP markers are genotyped for all patients, aimed at the identification of QTLs affecting virus dynamics. Suppose there is a functional QTL of alleles A and a for virus dynamics. Let q and $1-q$ denote the allele frequencies of A and a. The QTL forms three possible genotypes, AA, Aa, and aa. We assume that this QTL is associated with a SNP marker of alleles M (in a frequency of p) and m (in a frequency of $1-p$). The detection of significant linkage disequilibrium between the marker and QTL implies that the QTL may be linked with and,therefore, can be genetically manipulated by the marker.

The four haplotypes for the marker and QTL are MA, Ma, mA and ma, with respective frequencies expressed as $p_{11} = pq+D$, $p_{10} = p(1-q)-D$, $p_{01} = (1-p)q-D$, and $p_{00} = (1-p)(1-q)+D$, where D is the linkage disequilibrium between the marker and QTL. Thus, the population genetic parameters (p, q, and D) can be estimated by solving a group of regular equations if we can estimate the four haplotype frequencies $\mathbf{\Phi} = (p_{11}, p_{10}, p_{01}, p_{00})$. Joint marker-QTL diplotype frequencies can be expressed as a product of the corresponding haplotype frequencies under the HWE assumption, from which joint marker-QTL genotype frequencies are derived. Because the marker is observed, an unknown genotype of the QTL can be inferred from the conditional probability of the QTL genotype given a marker genotype.

Each sampled patient is measured for three different traits, uninfected cells, x, infected cells, y, and free virus particles, v, at a series of time points, $(t_{i1}, \ldots, t_{iT_i})$.

2.3.2 Likelihood

For a given QTL genotype j (j = 2 for AA, 1 for Aa, or 0 for aa), the parameters describing virus dynamics are denoted by $\mathbf{\Theta}_j = \{\lambda_j, d_j, \beta_j, a_j, k_j, u_j\}$. The comparisons of these parameters between the three different QTL genotypes can determine whether and how this QTL affects the pattern of virus dynamics.

The likelihood of longitudinal viral data $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i) = \{x_i(t_k), y_i(t_k), v_i(t_k)\}_{k=0}^{N}$ and marker information $\mathbf{M}_i$ for patient i is formulated by the mixture transitional Markov model, expressed as

$$L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = \prod_{i=1}^{n} \left[\sum_{j=0}^{2} \omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \mathbf{\Theta}_j, \mathbf{\Psi}) \right], \tag{2–5}$$

where $\omega_{j|i}$ is a mixture proportion, that is, the conditional probability of QTL genotype j given the marker genotype of subject i, which can be expressed as a function of haplotype frequencies (Table 1), and $f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \mathbf{\Theta}_j, \mathbf{\Psi})$ is a multivariate normal distribution with QTL genotype-specific mean vector specified by ODE parameters $(\mathbf{\Theta}_j)$ and covariance matrix specified by parametric, nonparametric, or semiparametric models $(\mathbf{\Psi})$ [10,15].

Based on the **Corollaries** given above, the multivariate distribution can be specified by the following transition model

$$f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \mathbf{\Theta}_j, \mathbf{\Psi}) = f_j(x_{i1}, y_{i1}, v_{i1} | \mathbf{\Theta}_j, \mathbf{\Psi}) \prod_{k=0}^{N-1} f_j(x_{ik+1}, y_{ik+1}, v_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \mathbf{\Psi}) \tag{2–6}$$

where

$$\begin{aligned}
& f_j(x_{ik+1}, y_{ik+1}, v_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \mathbf{\Psi}) \\
= \; & f_j(x_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \mathbf{\Psi}) f_j(y_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \mathbf{\Psi}) f_j(v_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \mathbf{\Psi}),
\end{aligned}$$

$$f_j(x_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \sigma_x^2) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{1}{2\sigma_x^2} \left(x_{ik+1} - g_j(x_{ik+1}) \right)^2 \right],$$

$$f_j(y_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \sigma_y^2) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[-\frac{1}{2\sigma_y^2} \left(y_{ik+1} - h_j(y_{ik+1}) \right)^2 \right],$$

$$f_j(v_{ik+1} | x_{ik}, y_{ik}, v_{ik}; \mathbf{\Theta}_j, \sigma_v^2) = \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left[-\frac{1}{2\sigma_v^2} \left(v_{ik+1} - l_j(v_{ik+1}) \right)^2 \right],$$

with $\boldsymbol{\Psi} = (\sigma_x^2, \sigma_y^2, \sigma_v^2)$, and

$$\begin{aligned} g_j(x_{ik+1}) &= x_{ik} + \lambda_j \Delta t_k - d_j x_{ik} \Delta t_k - \beta_j x_{ik} v_{ik} \Delta t_k \\ h_j(y_{ik+1}) &= y_{ik} + \beta_j x_{ik} v_{ik} \Delta t_k - a_j y_{ik} \Delta t_k \\ l_j(v_{ik+1}) &= v_{ik} + k_j y_{ik} \Delta t_k - u_j v_{ik} \Delta t_k. \end{aligned} \tag{2–7}$$

2.3.3 Estimation and Algorithm

The EM algorithm [18,19] is implemented to get the maximum likelihood estimates (MLE) of all unknown parameters. The gradient of the log-likelihood function

$$\log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = \sum_{i=1}^{n} \log \left[\sum_{j=0}^{2} \omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi}) \right], \tag{2–8}$$

is given by

$$\bigtriangledown_{\Theta_j} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})}{\sum_{j\prime=0}^{2} \omega_{j\prime|i} f_{j\prime}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})} \bigtriangledown_{\Theta_j} \log f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi}),$$

$$\bigtriangledown_{\Psi} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})}{\sum_{j\prime=0}^{2} \omega_{j\prime|i} f_{j\prime}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})} \bigtriangledown_{\Psi} \log f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi}),$$

and

$$\bigtriangledown_{\omega_{j|i}} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})}{\sum_{j\prime=0}^{2} \omega_{j\prime|i} f_{j\prime}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})} \bigtriangledown_{\omega_{j|i}} \log(\omega_{j|i}).$$

A iterative loop for the EM algorithm is formulated as follows. In the E step, the posterior probability with which a patient i carries a specific QTL genotype j based on the marker and phenotypic data is calculated by

$$\Omega_{j|i} = \frac{\omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})}{\sum_{j\prime=0}^{2} \omega_{j\prime|i} f_{j\prime}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})}. \tag{2–9}$$

In the M step, the parameters are estimated by solving the following log-likelihood equations:

$$\bigtriangledown_{\Theta_j} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = 0, \tag{2–10}$$

$$\bigtriangledown_{\Psi} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = 0, \tag{2–11}$$

$$\bigtriangledown_{\omega_{j|i}} \log L(\mathbf{x}, \mathbf{y}, \mathbf{v}; \mathbf{M}) = 0. \tag{2–12}$$

Wang and Wu [11] proposed a closed algorithmic form to obtain the MLEs of haplotype frequencies p_{11}, p_{10}, p_{01} and p_{00} and, therefore, allele frequencies of the marker (p) and QTL (q) and their linkage disequilibrium (D) without proof. These will be proved in Chapter 4. Genotype-specific mathematical parameters for viral dynamics and variances for the three types of viruses are calculated by implementing the Newton algorithm with the Armijo search [3].

2.4 Hypothesis Testing

2.4.1 The Significance of QTL

Whether there is a specific QTL responsible for viral dynamics described by a system of differential equations (2–1) can be tested by using the following hypotheses:

$$\begin{aligned} H_0 &: \ \boldsymbol{\Theta}_j \equiv \boldsymbol{\Theta}, \ (j = 2, 1, 0) \\ H_1 &: \ \text{At least one of the equalities above does not hold,} \end{aligned} \tag{2–13}$$

The likelihoods under the null (L_0) and alternative hypotheses (L_1) are calculated, from which a log-likelihood ratio test statistic is computed by

$$\text{LR} = -2[(\log L_0(\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}|\mathbf{x}, \mathbf{y}, \mathbf{z}) - \log L_1(\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Theta}}_j, \hat{\boldsymbol{\Psi}}|\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{M})],$$

where the tildes and hats present the maximum likelihood estimates under the null and alternative hypotheses, respectively. Because of violation of the regularity assumption, the LR may not asymptotically follow a χ^2-distribution with the degrees of freedom equal to the difference of parameter numbers between the two hypotheses (2–13). For this

reason, the threshold for claiming the existence of a significant QTL is determined from empirical permutation tests [21] because this approach does not rely on the distribution of LR values.

After a significant QTL is claimed, its significant association with the marker considered can be tested by the following hypotheses:

$$H_0 : D = 0 \text{ vs. } H_1 : D \neq 0, \tag{2–14}$$

whose log-likelihood ratio test statistic asymptotically follow the χ^2-distribution with one degree of freedom.

2.4.2 Genetic Mechanisms

The model allows the test of whether the QTL triggers a pleiotropic effect on three different types of cells. To do so, three null hypotheses for uninfected cells, infected cells, and free virus particles are formulated as follows:

$$H_0 : \quad (\lambda_j, d_j, \beta_j) \equiv (\lambda, d, \beta), \tag{2–15}$$

$$H_0 : \quad (\beta_j, a_j) \equiv (\beta, a), \tag{2–16}$$

$$H_0 : \quad (k_j, u_j) \equiv (k, u), \tag{2–17}$$

for $j = 2, 1, 0$. If all the null hypotheses are rejected, then this means that the QTL pleiotropically affect these three different aspects of viral dynamics. The pleiotropic effect of the QTL on any pair of three types of cells can also be tested accordingly. An empirical approach for determining the critical threshold is based on simulation studies.

2.4.3 Physiological Control of QTL

Several physiological important parameters define the dynamic system (2–1), including

(1) The average life-times, $1/d$, $1/a$, and $1/u$, of uninfected cells, infected cells, and free virus, respectively,

(2) The average number of virus particles or the burst size, k/a, yielded over the lifetime of a single infected cell,

(3) Basic reproductive ratio, $R_0 = \beta\lambda k/(adu)$, i.e., the average number of newly infected cells that arise from any one infected cell when almost all cells are uninfected.

How a QTL affects these physiological aspects of viral dynamics separately or jointly can be tested.

2.5 Application to Simulated Data

Monte Carlo simulation was performed to examine the statistical properties of the model for genetic mapping of viral dynamics. Also, the use of the model to analyze simulated data will validate its practical usefulness and utilization. We randomly choose 100 subjects from an HWE population. Consider one of the markers genotyped for all subjects. This marker of two alleles M and m is used to infer a QTL of two alleles A and a for viral dynamics based on the non-random association between the marker and QTL. The allele frequencies are assumed as $p = 0.6$ for allele M 0.4 for allele m as well as $q = 0.6$ for allele A and 0.4 for allele a. A positive value of linkage disequilibrium ($D = 0.08$) between alleles M and A is assumed, suggesting that these two more common alleles are in coupling phase.

The three QTL genotypes, AA, Aa, and aa, are each hypothesized to have different response systems for uninfected cells, x, infected cells, y, and free virus particles, v, constructed by equations (1). Six curve parameters $\{\lambda_j, d_j, \beta_j, a_j, k_j, u_j\}$ that define QTL genotype-specific systems were chosen from their spaces of biological relevance [8]. The phenotypic values of these three variables are expressed as the sum of the genotype-specific means and innovation errors assumed to follow a multivariate normal distribution. The phenotypic data were simulated for a practically reasonable number of equally spaced time points (say 22) under two different levels of heritability, low (0.1) and high (0.4). The genetic variance due to the QTL for virus response at a middle measurement point was used to define the heritability. The residual variances

for each of the three virus traits were then calculated under different heritabilities. To assure the homoscedasticity of variances, the transforms-both-sides (TBS) model was used to simulate innovation errors. The TBS model can preserve biological means of parameters in original differential equations and also avoid negative phenotypic values [12].

The differential equation-incorporated functional mapping model was used to analyze the simulated data, with the results suggesting that the QTL responsible for the dynamic system of viral infection can be detected using a molecular marker in association with the QTL. As expected, population genetic parameters about QTL segregation in a population can well be estimated with a closed form of the EM algorithm derived in [11]. The curve parameters for virus responses of each QTL genotype can be estimated accurately and precisely with a modest sample size (100) even for a low heritability of viral loads (Tables 1.2 and 1.3). The precision of all parameters can increase with increasing heritability level. By drawing the curves of viral trajectories with six parameters, the dynamic behavior of the system can be visualized. Figure 1 illustrates QTL genotype-specific curves of uninfected cells, infected cells, and free virus particles in a dynamic system from a random run of simulation. It is found that the shapes of the estimated curves are broadly consistent with the those of the true curves, suggesting that the system can be reasonably estimated with the new model.

Simulation studies showed that the new model displays reasonably high power, 0.75 for a modest heritability (0.1) and 0.99 for a high heritability (0.4), to detect a significant QTL responsible for a dynamic system of viral infection. Hypothesis tests described in Sections 1.4.2 and 1.4.3 provide a general platform for addressing the genetic control machinery of viral dynamics. For a given set of simulation data, it appears that these tests can be reasonably made. For example, the power for detecting a pleiotropic QTL for three types of viral cell dynamics is adequately high (≥ 0.7) for a modest sample size

and heritability level. On the other hand, under this circumstance, type I error rates for detecting a significant QTL despite its absence is reasonably low (≤ 0.1). These results suggest that our model will be practically useful in statistical analysis of the genetic control of viral dynamics.

2.6 Discussion

A combination of functional mapping [10–15] and mathematical models [7-9] provides new insights into the genetic control of virus population dynamics. In this article, we have proposed a statistical model for mapping quantitative trait loci (QTLs) that affect the dynamic pattern of viral infection. One of the meritorious advantages of the new model, as compared to existing functional mapping models, lies in the organization of multiple correlated aspects of viral infection into a dynamic system through a group of ordinary differential equations and the implementation of such a viral dynamic system into the framework of functional mapping. To our best knowledge, the work presented here is a first model of genetic mapping which treats multiple complex traits as a complex system.

The current model is not a simple extension of functional mapping for multiple traits [22]. The previous multi-trait models do not take into account the relationships of genotypic values of different traits, although they model across-trait correlations in residual errors. The new model views multiple traits as a whole in which different traits coordinate each other to determine the dynamic behavior of the system. Thus, by altering one variable or trait, other variables will change, leading to the change of the entire system. The genetic mapping of genes for a dynamic system will provide a powerful means for understanding the genetic architecture of a biological process.

The mathematical strength of the new model is the deployment of a system of differential equations in a genetic mapping context. The solution of multiple differential equations, especially high-dimensional ones, is computationally challenging. In this article, we apply a Newton algorithm with the EM setting to provide numerical estimates

of the parameters that define the dynamic system. With the corollaries derived from several assumptions of independence, the algorithm is shown from simulation studies to be computationally efficient and provides precise estimates of the parameters, even when the sample size used is modest (Tables 2-2 and 2-3).

As a demonstration of the new model, we assume that a dynamic system is controlled by a single QTL, although this assumption is too simple in real world. The genome-wide modeling of multiple QTLs throughout the genome can be incorporated into the current model setting, allowing the characterization of epistatic interactions among different QTLs [23,24]. A multi-locus linkage disequilibrium model has been available to specify high-order non-random associations among multiple loci in a natural population [25]. Although more parameters are involved in a multi-locus model, the closed forms derived for the EM algorithm [25] facilitates the estimation of many parameters at the same time. Also, a multi-locus model allows the test of the role of genetic interference in recombination events between adjacent intervals. Although linkage disequilibrium mapping has proven to be powerful for the high-resolution of QTLs, it often gives spurious results due to population structure and other evolutionary forces. A new genetic design that samples a set of random families, each composed of parents and their offspring, can overcome this limitation of linkage disequilibrium mapping [25,26]. This design allows the simultaneous estimation of the linkage and linkage disequilibrium between different genes, thus making it possible to construct a genome-wide linkage disequilibrium map for gene discovery.

Our model focuses on the identification of genes for a dynamic system of viral changes in a host's body before the administration of an anti-viral drug. When the patients are treated with a drug, the equilibrium state of the system will be violated, from which a new equilibrium will be generated. Bonhoeffer et al. [8] described a series of differential equations that specify the dynamic change of the system after drug treatment. The current model can be readily extended to model the genetic control of

Table 2-1. Joint genotype frequencies at the marker and QTL in terms of gametic haplotype frequencies, from which the conditional probabilities of QTL genotypes given marker genotypes can be calculated according to Bayes' theorem.

		AA	Aa	aa	
Genotype	Diplotype	$A\|A$	$A\|a + a\|A$	$a\|a$	Observations
MM	$M\|M$	p_{11}^2	$2p_{11}p_{10}$	p_{10}^2	N_1
Mm	$M\|m$	$2p_{11}p_{01}$	$2p_{11}p_{00} + 2p_{10}p_{01}$	$2p_{10}p_{00}$	N_2
mm	$m\|m$	p_{01}^2	$2p_{01}p_{00}$	p_{00}^2	N_3

viral declines in a response to the anti-viral drug and half-lives of infected cells in the body. Perhaps, the most promising aspect of the new model is that it, when incorporated with the dynamics of virus' drug resistance, can provide scientific guidance for drug delivery and development by characterizing genes for drug resistance. The emergence of drug-resistant virus presents a main problem with antiviral therapy. A system of differential equations that captures the essential dynamics of resistance is given in the literature [27,28]. With the idea presented in this Chapter, they can be readily incorporated into the functional mapping model, in a hope to achieve the maximum prevention of virus resistance to drugs by determining an optimal administration dose and time for individual patients based on their genetic makeups.

Table 2-2. The MLEs of parameters that define virus-host dynamics for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Virus-Host Parameters						
	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
λ	11.00	11.9556(0.0137)	10.00	10.99626(0.0115)	12.20	012.6963(0.0142)
d	0.4500	0.4622(0.0017)	0.090	0.1042(0.0014)	0.3800	0.4004(0.0017)
β	0.1100	0.1196(0.0001)	0.1200	0.1296(0.0001)	0.1300	0.1435(0.0031)
a	0.2000	0.1971(0.0008)	0.2500	0.2477(0.0006)	0.2500	0.2654(0.0008)
k	0.3000	0.3089(0.0010)	0.2000	0.2121(0.0008)	0.2100	0.2187(0.0012)
u	0.9800	1.0034(0.0025)	0.6400	0.6621(0.0023)	0.6800	0.7043(0.0033)

Genetic Parameters and Variances		
	Given	MLE
p	0.6	0.6010(0.0312)
q	0.6	0.5703(0.0680)
D	0.08	0.07701(0.0108)
σ_x^2	0.2538	0.2548(0.0107)
σ_y^2	0.2583	0.2978(0.0105)
σ_v^2	0.5976	0.5968(0.0239)

Table 2-3. The MLEs of parameters that define virus-host dynamics for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.4$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Virus Host Parameters						
	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
λ	11.000	11.9507(0.0069)	10.00	10.9599(0.0057)	12.00	12.4235(0.0100)
d	0.4500	0.4629(0.0009)	0.090	0.1046(0.0007)	0.380	0.3971(0.0012)
β	0.1100	0.1198(0.0000)	0.120	0.1278(0.0000)	0.130	0.1386(0.0000)
a	0.2000	0.1976(0.0003)	0.250	0.2480(0.0003)	0.250	0.2642(0.0005)
k	0.3000	0.3061(0.0004)	0.200	0.2106(0.0003)	0.210	0.2192(0.0006)
u	0.9800	1.0013(0.0011)	0.640	0.6626(0.0009)	0.680	0.7025(0.0017)

Genetic Parameters		
	Given	MLE
p	0.6	0.5976(0.0336)
q	0.6	0.6030(0.0407)
D	0.08	0.0793(0.0047)
σ_x^2	0.0423	0.0456(0.0020)
σ_y^2	0.0430	0.0436(0.0031)
σ_v^2	0.0996	0.1000(0.0045)

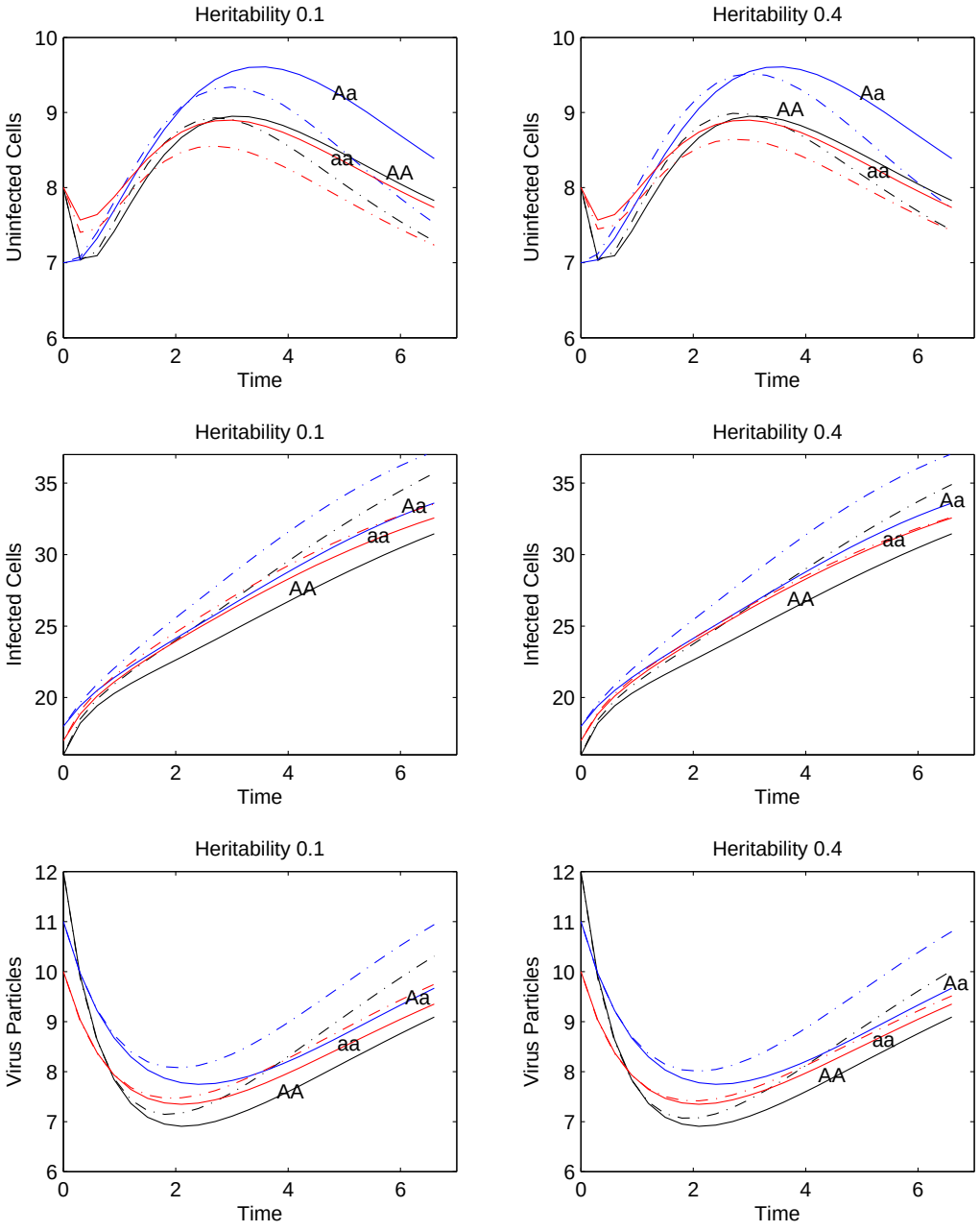

Figure 2-1. Estimated and true curves for a system of viral infection including uninfected cells, x, infected cells, y, and free virus particles, v for three genotypes at a simulated QTL, AA, Aa, and aa, under different heritability levels, 0.1 (right panel) and 0.4 (left panel). The broad consistency between the estimated and true curves suggests that the model can provide a reasonably good estimate of the dynamic system.

CHAPTER 3
BAYESIAN INFERENCE FOR GENETIC MAPPINGS OF DRUG RESISTANCE

3.1 Introduction

Antiretroviral drugs designed to treat infection by retroviruses, primarily HIV, act by inhibiting specific steps in the viral replication cycle [4, 6, 88]. Reverse-transcriptase inhibitors prevent the reverse transcription of viral genomic RNA into proviral DNA and thereby prevent the infection of new cells. Protease inhibitors influence the cleavage of viral polyproteins, resulting in the production of noninfectious virus particles. Both types of drugs are effective in reducing the viral load of infected individuals. Before treatment is initiated, the virus load in the individual's body is in a quasi steady state, i.e., the virus load is constant over a short period of time [4]. When a drug is administered, the virus load declines dramatically over several orders of magnitude after experiencing a transient shoulder phase. However, if monotherapy is used, resistant virus will rebound rapidly, in some cases within only a few weeks after the start of therapy.

While mathematical models have been widely used to study the decline of free virus in treated patients [35, 77, 80, 81], a growing body of interest has emerged in modeling the dynamics of viral drug resistance using a system of differential equations [4–7, 88, 90, 91]. Mathematical models may be instrumental in shedding some light on the prediction of the emergence of drug-resistant virus and ultimately the design of long-term therapy. Given a great deal of variation in the rate and pattern of rebound of resistant virus among different hosts [98], there is a pressing demand on the integration of genetic information into mathematical models for the precise prediction of the dynamics of decline in virus load during drug therapy and the rate of emergence of resistant virus. The increasing availability of single nucleotide polymorphism (SNP) data has made it possible to characterize concrete nucleotides or their combinations that encode a complex phenotype and, ultimately, document, map and understand

the structure and patterns of the human genome linked to the trait (The International HapMap Consortium 2003[34]).

The purpose of this chapter is to use Bayesian to derive a statistical model for identifying QTLs responsible for viral drug resistance. The new model is founded on a set of random samples drawn from a natural population at Hardy-Weinberg equilibrium. The Bayesian Markov chain properties of dynamic data will be incorporated into the model to facilitate the estimation of parameters that define virus dynamics. We perform simulation studies to investigate statistical properties of the model and validate its usefulness and utilization. For standard Bayesian statistics theory, please see [89]. For data analysis and modeling, our reference is [30].

3.2 Dynamic Models of Drug Resistance

3.2.1 Differential Equations

The basic model for describing virus dynamics is 2–1. The emergence of drug resistant virus in the therapy can be described by incorporating the difference of wild-type and mutant viruses into the equations, which is expressed as

$$\begin{aligned}
\frac{dx}{dt} &= \lambda - dx - \beta_1 x v_1 - \beta_2 x v_2 \\
\frac{dy_1}{dt} &= \beta_1(1-\mu) x v_1 + \beta_2 \mu x v_2 - a y_1 \\
\frac{dy_2}{dt} &= \beta_1 \mu x v_1 + \beta_2(1-\mu) x v_2 - a y_2 \\
\frac{dv_1}{dt} &= k_1 y_1 - u v_1 \\
\frac{dv_2}{dt} &= k_2 y_2 - u v_2,
\end{aligned} \tag{3–1}$$

where there are five variables: uninfected cells, x, cells infected by wild-type virus, y_1, cells infected by mutant virus, y_2, free wild-type virus, v_1, and free mutant virus, v_2. These five types of cells interact with each other to determine the dynamic changes of drug resistant virus in a host's body.

The system (3–1) is defined by nine parameters $\{\lambda, d, \beta_1, \beta_2, \mu, a, k_1, k_2, u\}$ and the initial conditions for x, y_1, y_2, v_1, and v_2. The dynamic pattern of this system can be determined and predicted by the change of these parameters and the initial conditions of x, y_1, y_2, v_1, and v_2. Again the application force us to introduce the Markov model by using Euler scheme

$$\begin{aligned}
\frac{x(t_{k+1}) - x(t_k)}{\Delta t_k} &= \lambda - dx(t_k) - \beta_1 x(t_k) v_1(t_k) - \beta_2 x(t_k) v_2(t_k) \\
\frac{y_1(t_{k+1}) - y_1(t_k)}{\Delta t_k} &= \beta_1(1-\mu) x(t_k) v_1(t_k) + \beta_2 \mu x(t_k) v_2(t_k) - a y_1(t_k) \\
\frac{y_2(t_{k+1}) - y_2(t_k)}{\Delta t_k} &= \beta_1 \mu x(t_k) v_1(t_k) + \beta_2(1-\mu) x(t_k) v_2(t_k) - a y_2(t_k) \\
\frac{v_1(t_{k+1}) - v_1(t_k)}{\Delta t_k} &= k_1 y_1(t_k) - u v_1(t_k) \\
\frac{v_2(t_{k+1}) - v_2(t_k)}{\Delta t_k} &= k_2 y_2(t_k) - u v_2(t_k).
\end{aligned} \tag{3–2}$$

For long term treatment not only the the drug resistance appears but also the parameters change drastically. The longitudinal observations do not follow the model introduced in Chapter 1. A better model is to assume that the parameters themselves are random observations following certain distributions.

3.2.2 Bayesian Markov Model for Drug Resistance

A Bayesian transitional Markov model is used to describe the random process of the drug resistance system by

$$\begin{aligned}
x_i(t_{k+1}) &= x_i(t_k) + \lambda \Delta t_k - d x_i(t_k) \Delta t_k - \beta_1 x_i(t_k) v_{1i}(t_k) \Delta t_k - \beta_2 x_i(t_k) v_{2i}(t_k) \Delta t_k + \epsilon_{x_i}(t_k) \\
y_{1i}(t_{k+1}) &= y_{1i}(t_k) + \beta_1(1-\mu) x_i(t_k) v_{1i}(t_k) \Delta t_k + \beta_2 \mu x_i(t_k) v_2(t_k) \Delta t_k - a y_1(t_k) \Delta t_k + \epsilon_{y_{1i}}(t_k) \\
y_{2i}(t_{k+1}) &= y_{2i}(t_k) + (\beta_1 \mu x_i(t_k) v_{1i}(t_k) + \beta_2(1-\mu) x_i(t_k) v_{2i}(t_k) - a y_{2i}(t_k)) \Delta t_k + \epsilon_{y_{2i}}(t_k) \\
v_{1i}(t_{k+1}) &= v_{1i}(t_k) + k_1 y_{1i}(t_k) \Delta t_k - u v_{1i}(t_k) \Delta t_k + \epsilon_{v_{1i}}(t_k) \\
v_{2i}(t_{k+1}) &= v_{2i}(t_k) + k_2 y_{2i}(t_k) \Delta t_k - u v_{2i}(t_k) \Delta t_k + \epsilon_{v_{2i}}(t_k),
\end{aligned} \tag{3–3}$$

and the prior distributions

$$\begin{aligned}
\lambda &\sim N(\theta_\lambda, \sigma_\lambda^2) \\
d &\sim N(\theta_d, \sigma_d^2) \\
\beta_1 &\sim N(\theta_{\beta_1}, \sigma_{\beta_1}^2) \\
\beta_2 &\sim N(\theta_{\beta_2}, \sigma_{\beta_2}^2) \\
\mu &\sim N(\theta_\mu, \sigma_\mu^2) \\
a &\sim N(\theta_a, \sigma_a^2) \\
k_1 &\sim N(\theta_{k_1}, \sigma_{k_1}^2) \\
k_2 &\sim N(\theta_{k_2}, \sigma_{k_2}^2) \\
u &\sim N(\theta_u, \sigma_u^2),
\end{aligned} \tag{3–4}$$

where $\epsilon_{x_i}(t_k) \sim N(0, \sigma_x^2)$, $\epsilon_{y_{1i}}(t_k) \sim N(0, \sigma_{y1}^2)$, $\epsilon_{y_{2i}}(t_k) \sim N(0, \sigma_{y2}^2)$, $\epsilon_{v_{1i}}(t_k) \sim N(0, \sigma_{v1}^2)$, and $\epsilon_{v_{2i}}(t_k) \sim N(0, \sigma_{v2}^2)$ are the errors for five variables, x, y_1, y_2, v_1, and v_2, respectively, each of which is assumed to be iid and time-independent. Again we assume that these error terms of the five variables are independent of each other, although this assumption can be relaxed. The priors are also assumed to be independent, which is biologically meaningful.

For simplicity, we use x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, and v_{2ik} to stand for $x_i(t_k)$, $y_{1i}(t_k)$, $y_{2i}(t_k)$, $v_{1i}(t_k)$, and $v_{2i}(t_k)$, respectively. Just as in Chapter 1, we have:

Theorem 2.2.1: *All the future values of uninfected cells, cells infected by wild-type virus, cells infected by mutant virus, free wild-type virus, and free mutant virus depend*

statistically only on their present values, i.e.,

$$f(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{2ik+1}|(x_{i1}, y_{1i1}, y_{2i1}, v_{1i1}, v_{2i1}), \dots, (x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}))$$

$$= f(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{ik+1}|(x_{ik}, y_{ik}, v_{ik})),$$

$$f(x_{ik+1}|(x_{i1}, y_{1i1}, y_{2i1}, v_{1i1}, v_{2i1}), \dots, (x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}))$$

$$= f(x_{ik+1}|(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})),$$

$$f(y_{ik+1}|(x_{i1}, y_{1i1}, y_{2i1}, v_{1i1}, v_{2i1}), \dots, (x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{1ik}))$$

$$= f(y_{ik+1}|(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})),$$

$$f(v_{ik+1}|(x_{i1}, y_{1i1}, y_{2i1}, v_{1i1}, v_{i1}), \dots, (x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{ik}))$$

$$= f(v_{ik+1}|(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{ik})).$$

Furthermore, from this theorem, we have the following results.

Corollary 2.2.1 *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $(x_{ik-1}, y_{1ik-1}, y_{2ik-1}, v_{1ik-1}, v_{2ik-1})$ *and* $(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{ik+1})$ *are statistically independent.*

Corollary 2.2.2: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, x_{ik-1} *and* x_{ik+1} *are statistically independent.*

Corollary 2.2.3: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, y_{1ik-1} *and* y_{1ik+1} *are statistically independent.*

Corollary 2.2.4: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, y_{2ik-1} *and* y_{2ik+1} *are statistically independent.*

Corollary 2.2.5: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, v_{1ik-1} *and* v_{1ik+1} *are statistically independent.*

Corollary 2.2.6: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, v_{2ik-1} *and* v_{2ik+1} *are statistically independent.*

Corollary 2.2.7: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $(x_{ij}, y_{1ij}, y_{2ij}, v_{1ij}, v_{2ij})$ *for* $j = 0, 1, \dots, k-1$ *and* $(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{2ik+1})$ *are statistically independent.*

Corollary 2.2.8: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $\{x_{i1}, \ldots, x_{ik-1}\}$, *and* x_{ik+1} *are statistically independent.*

Corollary 2.2.9: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $\{y_{1i1}, \ldots, y_{1ik-1}\}$, *and* y_{1ik+1} *are statistically independent.*

Corollary 2.2.10: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $\{v_{1i1}, \ldots, v_{1ik-1}\}$, *and* v_{1ik+1} *are statistically independent.*

Corollary 2.2.11: *Conditional on* $(x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik})$, $\{v_{2i1}, \ldots, v_{2ik-1}\}$, *and* v_{2ik+1} *are statistically independent.*

Of course all the above results condition on the random parameters and the proof of them is similar to that of Chapter 1.

3.3 Genetic Mapping for Drug Resistance

3.3.1 Likelihood Function

We use the same genetic design as Chapter 1. For a patient the dynamic model is described by the random vector $\boldsymbol{\Upsilon}_i = \{\lambda_i, d_i, \beta_{1i}, \beta_{2i}, \mu_i, a_i, k_{1i}, k_{2i}, u_i\}$ of prior distribution. The $\boldsymbol{\Upsilon}_i$ for the patients with the same QTL genotype j have the same distribution with mean $\{\theta_{j\lambda}, \theta_{jd}, \theta_{j\beta_1}, \theta_{j\beta_2}, \theta_{j\mu}, \theta_{ja}, \theta_{jk_1}, \theta_{jk_2}, \theta_{ju}\}$ (j = 2 for AA, 1 for Aa, or 0 for aa). The drug resistance is determined by the distribution and structure of these random vectors.

The likelihood of longitudinal viral data $(\mathbf{x}_i, \mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{v}_{1i}, \mathbf{v}_{2i}) = \{x_i(t_k), y_{1i}(t_k), y_{2i}(t_k), v_{1i}(t_k), v_{2i}(t_k)\}_{k=0}^{N}$ and marker information $\mathbf{M}_i$ for patient i is formulated by the mixture of Bayesian transitional Markov model, expressed as

$$L(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{v}_1, \mathbf{v}_2; \mathbf{M}) = \prod_{i=1}^{n} \left[\sum_{j=0}^{2} \omega_{j|i} f_j(\mathbf{x}_i, \mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{v}_{1i}, \mathbf{v}_{2i} | \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) f(\boldsymbol{\Upsilon}_j | \boldsymbol{\Theta}_j) \right], \qquad (3\text{–}5)$$

where $\omega_{j|i}$ is a mixture proportion that reflects the QTL genotype j of patient i, and $f_j(\mathbf{x}_i, \mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{v}_{1i}, \mathbf{v}_{2i}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi})$ is a multivariate normal distribution with QTL genotype-specific random vector specified by $\boldsymbol{\Upsilon}_j$ and covariance matrix specified by $\boldsymbol{\Psi}$.

The multivariate distribution can be specified by the following transition model

$$f_j(\mathbf{x}_i, \mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{v}_{1i}, \mathbf{v}_{2i}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) = f_j(x_{i1}, y_{1i1}, y_{2i1}, v_{1i1}, v_{2i1}|\boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) \times \prod_{k=1}^{N-1} f_j(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) \quad (3\text{–}6)$$

where

$$\begin{aligned} & f_j(x_{ik+1}, y_{1ik+1}, y_{2ik+1}, v_{1ik+1}, v_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) \\ = \; & f_j(x_{ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) f_j(y_{1ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) \\ & f_j(y_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) f_j(v_{1ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}) \\ & f_j(v_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \boldsymbol{\Psi}), \end{aligned}$$

$$f_j(x_{ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \sigma_x^2) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{1}{2\sigma_x^2}\left(x_{ik+1} - g_j(x_{ik})\right)^2\right],$$

$$f_j(y_{1ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \sigma_{y1}^2) = \frac{1}{\sqrt{2\pi\sigma_{y1}^2}} \exp\left[-\frac{1}{2\sigma_{y1}^2}\left(y_{1ik+1} - h_{1j}(y_{1ik})\right)^2\right],$$

$$f_j(y_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \sigma_{y2}^2) = \frac{1}{\sqrt{2\pi\sigma_{y2}^2}} \exp\left[-\frac{1}{2\sigma_{y2}^2}\left(y_{2ik+1} - h_{2j}(y_{2ik})\right)^2\right],$$

$$f_j(v_{1ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \sigma_{v1}^2) = \frac{1}{\sqrt{2\pi\sigma_{v1}^2}} \exp\left[-\frac{1}{2\sigma_{v1}^2}\left(v_{1ik+1} - l_{1j}(v_{1ik})\right)^2\right],$$

$$f_j(v_{2ik+1}|x_{ik}, y_{1ik}, y_{2ik}, v_{1ik}, v_{2ik}; \boldsymbol{\Upsilon}_j, \sigma_{v2}^2) = \frac{1}{\sqrt{2\pi\sigma_{v2}^2}} \exp\left[-\frac{1}{2\sigma_{v2}^2}\left(v_{2ik+1} - l_{2j}(v_{2ik})\right)^2\right],$$

$$\boldsymbol{\Psi} = (\sigma_x^2, \sigma_{y1}^2, \sigma_{y2}^2, \sigma_{v1}^2, \sigma_{v2}^2),$$

$$\begin{aligned} g_j(x_{ik}) &= x_{ik} + \lambda_j \Delta t_k - d_j x_{ik} \Delta t_k - \beta_{1j} x_{ik} v_{1ik} \Delta t_k - \beta_{2j} x_{ik} v_{2ik} \Delta t_k \\ h_{1j}(y_{1ik}) &= y_{1ik} + \beta_{1j}(1-\mu_j) x_{ik} v_{1ik} \Delta t_k + \beta_{2j} \mu_j x_{ik} v_{2ik} \Delta t_k - a_j y_{1ik} \Delta t_k \\ h_{2j}(y_{2ik}) &= y_{2ik} + \beta_{1j} \mu_j x_{ik} v_{1ik} \Delta t_k + \beta_{2j}(1-\mu_j) x_{ik} v_{2ik} \Delta t_k - a_j y_{2ik} \Delta t_k \\ l_{1j}(v_{1ik}) &= v_{1ik} + k_{1j} y_{1ik} \Delta t_k - u_j v_{1ik} \Delta t_k \\ l_{2j}(v_{2ik}) &= v_{2ik} + k_{2j} y_{2ik} \Delta t_k - u_j v_{2ik} \Delta t_k, \end{aligned} \quad (3\text{–}7)$$

and,

$$
\begin{aligned}
f(\boldsymbol{\Upsilon}_j|\boldsymbol{\Theta}_j) \;=\; & \frac{1}{\sqrt{2\pi\sigma_\lambda^2}}\exp\left[-\frac{1}{2\sigma_\lambda^2}\left(\lambda_j-\theta_{j\lambda}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_d^2}}\exp\left[-\frac{1}{2\sigma_d^2}\left(d_j-\theta_{jd}\right)^2\right] \\
& \frac{1}{\sqrt{2\pi\sigma_{\beta_1}^2}}\exp\left[-\frac{1}{2\sigma_{\beta_1}^2}\left(\beta_{1j}-\theta_{j\beta_1}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_{\beta_2}^2}}\exp\left[-\frac{1}{2\sigma_{\beta_2}^2}\left(\beta_{2j}-\theta_{j\beta_2}\right)^2\right] \\
& \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\exp\left[-\frac{1}{2\sigma_\mu^2}\left(\mu_j-\theta_{j\mu}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left[-\frac{1}{2\sigma_a^2}\left(a_j-\theta_{ja}\right)^2\right] \\
& \frac{1}{\sqrt{2\pi\sigma_{k_1}^2}}\exp\left[-\frac{1}{2\sigma_{k_1}^2}\left(k_{1j}-\theta_{jk_1}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_{k_2}^2}}\exp\left[-\frac{1}{2\sigma_{k_2}^2}\left(k_{2j}-\theta_{k_2}\right)^2\right] \\
& \frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left[-\frac{1}{2\sigma_u^2}\left(u_j-\theta_{ju}\right)^2\right],
\end{aligned}
$$

3.3.2 Estimation and Algorithm

To get the estimates we need to solve the following optimization problem

$$
argmax\sum_{i=1}^{n} log\left[\sum_{j=0}^{2}\omega_{j|i}f_j(\mathbf{x}_i,\mathbf{y}_{1i},\mathbf{y}_{2i},\mathbf{v}_{1i},\mathbf{v}_{2i}|\boldsymbol{\Upsilon}_i,\boldsymbol{\Psi})f(\boldsymbol{\Upsilon}_i|\boldsymbol{\Theta}_j)\right]
$$

which is untractable since the Υ_i are unobservable. But our main focus is on the estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_j$ an alternative way is to work on marginal likelihood function in stead of the joint one. Let

$$
f_j(\mathbf{x}_i,\mathbf{y}_{1i},\mathbf{y}_{2i},\mathbf{v}_{1i},\mathbf{v}_{2i}|\boldsymbol{\Psi},\boldsymbol{\Theta}_j)=\int f_j(\mathbf{x}_i,\mathbf{y}_{1i},\mathbf{y}_{2i},\mathbf{v}_{1i},\mathbf{v}_{2i}|\boldsymbol{\Upsilon}_i,\boldsymbol{\Psi})f(\boldsymbol{\Upsilon}_i|\boldsymbol{\Theta}_j)d\boldsymbol{\Upsilon}_i
$$

then we solve

$$
argmax\sum_{i=1}^{n} log\left[\sum_{j=0}^{2}\omega_{j|i}f_j(\mathbf{x}_i,\mathbf{y}_{1i},\mathbf{y}_{2i},\mathbf{v}_{1i},\mathbf{v}_{2i}|\boldsymbol{\Psi},\boldsymbol{\Theta}_j)\right]
$$

for the estimates of $\boldsymbol{\Psi}$, and $\boldsymbol{\Theta}_j$; which can be similarly obtained as in Chapter 1 and the details are given in Chapter 4.

Our next task here is about testing.

3.3.3 Testing the Significance of QTL and Drug Resistance

The significance of QTL for viral dynamics of differential equations (3–1) can be tested by using the following hypotheses:

$$\begin{aligned} H_0: &\quad \mathbf{\Theta}_j \equiv \mathbf{\Theta},\ (j = 2, 1, 0) \\ H_1: &\quad \text{At least one of the equalities above does not hold,} \end{aligned} \tag{3–8}$$

The likelihoods under each hypothesis are calculated from which a log-likelihood ratio test statistic is calculated.

After a significant QTL is claimed, its significant association with the marker considered can be tested by the following hypotheses:

$$H_0 : D = 0 \text{ vs. } H_1 : D \neq 0, \tag{3–9}$$

whose log-likelihood ratio test statistic asymptotically follow theχ^2-distribution with one degree of freedom.

The significance of drug resistance can be tested by:

$$H_0: \quad (\lambda_j, d_j, \beta_{1j}, \beta_{2j}) \equiv (\lambda, d, \beta_1, \beta_2) \quad \text{uninfectedcells} \tag{3–10}$$

$$H_0: \quad (\mu_j, \beta_{1j}, \beta_{2j}, a_j) \equiv (\mu, \beta_1, \beta_2, a) \quad \text{cells infected by wild and mutant virus} \tag{3–11}$$

$$H_0: \quad (k_{1j}, u_j) \equiv (k_1, u) \quad \text{free wild} - \text{type virus} \tag{3–12}$$

$$H_0: \quad (k_{2j}, u_j) \equiv (k_2, u) \quad \text{free mutant virus,} \tag{3–13}$$

for $j = 2, 1, 0$. If all the null hypotheses are rejected, then this means that the QTL pleiotropically affect these five different aspects of viral dynamics. The pleiotropic effect of the QTL on any pair of five types of cells can also be tested accordingly. An empirical approach for determining the critical threshold is based on simulation studies.

3.4 Computer Simulation and Discussion

We perform a similar Monte Carlo simulation to examine the statistical properties of the model for genetic mapping of the dynamics of viral drug resistance. The results

are shown in Tables 2.1 and 2.2. These results suggest that our model will be practically useful in statistical analysis of the genetic control of drug resistance in viral dynamics.

The resistance of virus to an antiviral drug is one of the most important reasons of the failure of drug treatment to achieve complete viral suppression [21] . The identification of genes that control the dynamic pattern of viral drug resistance will provide useful information to understand the emergence of drug resistance and better predict treatment outcomes. In this chapter, we have developed a statistical model for mapping quantitative trait loci (QTLs) that affect the rate and pattern of rebound of resistant virus after drug therapy. This model integrates a system of differential equations for the dynamic change of viral drug resistance into functional mapping developed by Wu and group [64, 113], from which a number of hypotheses about the interplay between genetic actions and viral dynamics can be formulated and addressed.

The current model is not a simple extension of functional mapping for multiple dynamic traits [135]. The previous multi-trait models do not take into account the relationships of genotypic values of different traits, although across-trait correlations due to residual errors are considered. The new model views multiple traits as a whole in which different traits coordinate each other to determine the dynamic behavior of the system. Thus, by altering one variable or trait, other variables will change, leading to the change of the entire system. The genetic mapping of genes for a dynamic system will provide a powerful means for understanding the genetic architecture of a biological process.

The mathematical strength of the new model is the deployment of a system of differential equations in a genetic mapping context. The solution of multiple differential equations, especially high-dimensional ones, is computationally challenging. In this article, we apply a Newton algorithm within the EM setting to provide numerical estimates of the parameters that define the dynamic system. With the theorems derived from several assumptions of independence, the algorithm is shown from simulation

studies to be computationally efficient and provides precise estimates of the parameters, even when the sample size used is modest (Tables 3-1 and 3-2).

As a demonstration of the new model, we assume that a dynamic system is controlled by a single QTL, although this assumption is too simple in real world. The genome-wide modeling of multiple QTLs throughout the genome can be incorporated into the current model setting, allowing the characterization of epistatic interactions among different QTLs. A multi-locus linkage disequilibrium model has been available to specify high-order non-random associations among multiple loci in a natural population [114]. Although linkage disequilibrium mapping has proven to be powerful for the high-resolution of QTLs, it often gives spurious results due to population structure and other evolutionary forces. A new genetic design that samples a set of random families, each composed of parents and their offspring, can overcome this limitation of linkage disequilibrium mapping [55]. This design allows the simultaneous estimation of the linkage and linkage disequilibrium between different genes, thus making it possible to construct a genome-wide linkage disequilibrium map for gene discovery.

The emergence of drug-resistant virus may be due to the preexistence of drug resistant strains before the initiation of therapy or the generation of resistant virus during the course of treatment. It is important to identify which process is more likely to be true, drug resistant virus preexists before the onset of therapy or they are produced by residual virus replication during the course of antiviral treatment, because each process requires different drug regimens to maximize the clinical benefits [7]. [90, 91] developed a mathematical model to investigate analytically the mechanisms underlying the emergence of drug-resistant variants during antiviral treatment. By incorporating this mathematical model into our functional mapping framework, it is possible to test whether there is a specific QTL that determines each of these two processes and how they can be predicted with genetic information of the QTL detected.

Table 3-1. The MLEs of parameters that define the dynamics of viral drug resistance for three different QTL genotypes and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.05$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Virus-Host Parameters

	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
θ_λ	33.000	35.1656(0.0168)	32.00	30.5237(0.0111)	36.00	33.9577(0.0149)
θ_d	0.1000	0.1186(0.0048)	0.300	0.3033(0.00031)	0.600	0.6634(0.0046)
θ_{β_1}	0.7000	0.6305(0.0003)	0.500	0.5094(0.0003)	0.900	0.8085(0.0005)
θ_{β_2}	0.5000	0.5515(0.0005)	0.600	0.5016(0.0003)	0.340	0.3171(0.0005)
θ_μ	0.3000	0.3143(0.0039)	0.190	0.1746(0.0042)	0.260	0.27454(0.0066)
θ_a	0.3500	0.3322(0.0019)	0.200	0.2587(0.0014)	0.300	0.3255(0.0019)
θ_{k_1}	0.3700	0.4167(0.0013)	0.280	0.2619(0.0009)	0.200	0.220120.0015)
θ_{k_2}	0.1980	0.1837(0.0017)	0.160	0.1782(0.0012)	0.138	0.1243(0.0018)
θ_u	0.0550	0.0576(0.0015)	0.060	0.0658(0.0011)	0.020	0.0210(0.0019)

Genetic Parameters

	Given	MLE
p	0.6	0.6021(0.0339)
q	0.6	0.5089(0.0447)
D	0.08	0.0358(0.0352)
σ_x^2	0.2299	0.2347(0.0106)
σ_{y1}^2	1.3566	1.3672(0.0591)
σ_{y2}^2	1.0944	1.1041(0.0450)
σ_{v1}^2	0.3382	0.3412(0.0138)
σ_{v2}^2	0.4332	0.4326(0.0180)

Table 3-2. The MLEs of parameters that define the dynamics of viral drug resistance for three different QTL genotypes, and the association between the marker and QTL in a natural population, assuming that the heritability of the simulated QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Virus-Host Parameters

	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
θ_λ	33.000	34.4361(0.0112)	32.00	32.9949(0.0096)	36.00	36.9951(0.0136)
θ_d	0.1000	0.1377(0.0034)	0.300	0.3207(0.00027)	0.600	0.6205(0.0039)
θ_{β_1}	0.7000	0.6843(0.0009)	0.500	0.5199(0.0003)	0.900	0.9297(0.0004)
θ_{β_2}	0.5000	0.5120(0.0004)	0.600	0.6199(0.0003)	0.340	0.3449(0.0005)
θ_μ	0.3000	0.3196(0.0020)	0.190	0.2023(0.0044)	0.260	0.2706(0.0071)
θ_a	0.3500	0.3547(0.0011)	0.200	0.2118(0.0010)	0.300	0.3130(0.0016)
θ_{k_1}	0.3700	0.3783(0.0008)	0.280	0.2901(0.0007)	0.200	0.2205(0.0011)
θ_{k_2}	0.1980	0.2097(0.0010)	0.160	0.1798(0.0009)	0.138	0.1414(0.0015)
θ_u	0.0550	0.0580(0.0010)	0.060	0.0700(0.0008)	0.02	0.0210(0.0015)

Genetic Parameters

	Given	MLE
p	0.6	0.6021(0.0339)
q	0.6	0.5529(0.0478)
D	0.08	0.0683(0.0311)
σ_x^2	0.1089	0.1141(0.0053)
σ_{y1}^2	0.6426	0.6425(0.0281)
σ_{y2}^2	0.5184	0.5191(0.0212)
σ_{v1}^2	0.1602	0.1607(0.0065)
σ_{v2}^2	0.2052	0.2047(0.0085)

CHAPTER 4
EM ALGORITHM FOR SOLVING MIXTURE MODELS IN COMPLEX GENETIC TRAITS

4.1 Introduction

The study of the mixture models has more than one hundred years. There are more than one thousand papers and articles and several monographs about it. The application spreads almost all scientific and engineering areas. Our main focus in this chapter is to study the mixture densities of the form of (2–5), which are widely used in genetics. We are not going to review the literatures about the mixture models since this has been done by many authors (see [24, 65, 66, 87] and their references). First, we are going to give a brief review of genetic theory and EM algorithm. Then we discuss the theoretical properties of the EM algorithm for mixture model (2–5). We also fit a group of real genetic data and use computer to do simulation study.

4.2 Algorithm

4.2.1 Genetic Design

Here we briefly review the concept of genetic design used in Chapter 1. The natural human population from which n individuals are sampled is assumed to be at Hardy-Weinberg equilibrium (HWE). A panel of SNP markers are genotyped for all subjects, aimed at the identification of QTLs affecting a special growth factor. Suppose there is a functional QTL of alleles A and a for the growth. Let q and $1-q$ denote the allele frequencies of A and a. The QTL forms three possible genotypes, AA (symbolized by 2), Aa (symbolized by 1), and aa (symbolized by 0). We assume that this QTL is associated with a SNP marker of alleles M (in a frequency of p) and m (in a frequency of $1-p$). The detection of significant linkage disequilibrium between the marker and QTL implies that the QTL may be linked with and, therefore, can be genetically manipulated by the marker.

The four haplotypes for the marker and QTL are MA, Ma, mA, and ma, with respective frequencies expressed as $p_{11} = pq + D$, $p_{10} = p(1-q) - D$, $p_{01} = (1-p)q - D$,

and $p_{00} = (1-p)(1-q) + D$, where the linkage disequilibrium between the marker and QTL. Thus, the population genetic parameters (p, q, and D) can be estimated by solving a group of regular equations if we can estimate the four haplotype frequencies. Joint marker-QTL diplotype frequencies can be expressed as a product of the corresponding haplotype frequencies under the HWE assumption, from which joint marker-QTL genotype frequencies are derived. Because the marker is observed, an unknown genotype of the QTL can be inferred from the conditional probability of the QTL genotype given a marker genotype.

4.2.2 Likelihood

For a given QTL genotype j ($j = 0, 1, 2$), the parameters describing biological growth are denoted by $\mathbf{\Theta}_j$. The comparisons of these parameters between the three different QTL genotypes can determine whether and how this QTL affects the pattern of the trajectory.

The likelihood of longitudinal genetic data $\mathbf{Y}_i$ for $i = 1, \ldots, n$, with

$$\mathbf{Y}_i\prime = (Y_i(1), \ldots, Y_i(m))$$

and marker information $\mathcal{M}_j$ for patient i is formulated by the mixture model, expressed as

$$L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}) = \prod_{k=0}^{2} \prod_{i=1}^{m_k} \left[\sum_{j=0}^{2} \omega_{j|k}(p, q, D, \mathcal{M}_k) f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j) \right] P(\mathcal{M}_k|p), \quad (4\text{–}1)$$

where $\omega_{j|k}(p, q, D, \mathcal{M}_k)$, a function of p, q, D, and $\mathcal{M}_k$, is a mixture proportion that reflects the QTL genotype j of individual i with marker k, which can be inferred from its marker genotype [106], and $f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)$ is usually a multivariate normal distribution with QTL genotype-specific mean vector being function of $\mathbf{\Theta}_j$, and covariance matrix specified by $\mathbf{\Psi}_j$.

Again, based on [106],

$$P(\mathcal{M}_2|p) = p^2,\ P(\mathcal{M}_1|p) = 2p(1-p),\ P(\mathcal{M}_0|p) = (1-p)^2;$$

and

$$\omega_{2|2} = \frac{(pq+D)^2}{p^2},\ \omega_{1|2} = \frac{2(pq+D)(p(1-q)-D)}{p^2},\ \omega_{0|2} = \frac{(p(1-q)-D)^2}{p^2};$$

$$\omega_{2|1} = \frac{2(pq+D)((1-p)q-D)}{2p(1-p)},\ \omega_{0|1} = \frac{2(p(1-q)-D)((1-p)(1-q)+D)}{2p(1-p)},$$

$$\omega_{1|1} = \frac{2(pq+D)((1-p)(1-q)+D)+2(p(1-q)-D)((1-p)q-D)}{2p(1-p)};$$

$$\omega_{2|0} = \frac{((1-p)q-D)^2}{(1-p)^2},\ \omega_{0|0} = \frac{((1-p)(1-q)+D)^2}{(1-p)^2};$$

$$\omega_{1|0} = \frac{2((1-p)q-D)((1-p)(1-q)+D)}{(1-p)^2}.$$

Plugging the above formulae in (4–1) and canceling all common factors, we know that our likelihood function consists of the product of the following three parts

$$\prod_{i=1}^{m_1}[(pq+D)^2 f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2,\boldsymbol{\Psi}_2) + 2(pq+D)(p(1-q)-D)f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1,\boldsymbol{\Psi}_1) + (p(1-q)-D)^2 f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0,\boldsymbol{\Psi}_0)], \tag{4–2}$$

$$\prod_{i=1}^{m_2}[\Pi_1 f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2,\boldsymbol{\Psi}_2) + \Pi_2 f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1,\boldsymbol{\Psi}_1) + \Pi_3 f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0,\boldsymbol{\Psi}_0)], \tag{4–3}$$

with

$$\Pi_1 = 2(pq+D)((1-p)q-D),$$

$$\Pi_2 = (2(pq+D)((1-p)(1-q)+D) + 2(p(1-q)-D)((1-p)q-D)),$$

$$\Pi_3 = 2(p(1-q)-D)((1-p)(1-q)+D).$$

and

$$\prod_{i=1}^{m_3}[((1-p)q-D)^2 f_2(\mathbf{Y}_i|\mathbf{\Theta}_2,\mathbf{\Psi}_2)+2((1-p)q-D)((1-p)(1-q)+D) \cdot f_1(\mathbf{Y}_i|\mathbf{\Theta}_1,\mathbf{\Psi}_1)+((1-p)(1-q)+D)^2 f_0(\mathbf{Y}_i|\mathbf{\Theta}_0,\mathbf{\Psi}_0)]. \quad (4\text{–}4)$$

Therefore our likelihood function (4–1) can be rewritten as

$$L(\mathbf{\Theta},\mathbf{\Psi},p,q,D|\mathbf{Y};\mathcal{M})=\prod_{k=0}^{2}\prod_{i=1}^{m_k}\left[\sum_{j=0}^{2}\varpi_{j|k}(p,q,D,\mathcal{M}_k)f_j(\mathbf{Y}_i|\mathbf{\Theta}_j,\mathbf{\Psi}_j)\right] \quad (4\text{–}5)$$

with

$$\varpi_{2|2}=(pq+D)^2,\ \varpi_{1|2}=2(pq+D)(p(1-q)-D),\ \varpi_{0|2}=(p(1-q)-D)^2;$$

$$\varpi_{2|1}=2(pq+D)((1-p)q-D),\ \varpi_{0|1}=2(p(1-q)-D)((1-p)(1-q)+D),$$

$$\varpi_{1|1}=2(pq+D)((1-p)(1-q)+D)+2(p(1-q)-D)((1-p)q-D);$$

$$\varpi_{2|0}=((1-p)q-D)^2,\ \ \varpi_{0|0}=((1-p)(1-q)+D)^2;$$

$$\varpi_{1|0}=2((1-p)q-D)((1-p)(1-q)+D).$$

Or, simply

$$L(\mathbf{\Theta},\mathbf{\Psi},p,q,D|\mathbf{Y};\mathcal{M})=\prod_{i=1}^{n}\left[\sum_{j=0}^{2}\varpi_{j|i}(p,q,D,\mathcal{M}_k)f_j(\mathbf{Y}_i|\mathbf{\Theta}_j,\mathbf{\Psi}_j)\right] \quad (4\text{–}6)$$

Now our purpose is to maximize (4–6) with respect to p, q, D, $\mathbf{\Theta}$, and $\mathbf{\Psi}$ under the condition that f_j has multiple normal distribution or Markov model construction with normal distribution.

4.2.3 Algorithm

The gradient of log-likelihood function

$$logL(\mathbf{\Theta},\mathbf{\Psi},p,q,D|\mathbf{Y};\mathcal{M})=\sum_{i=1}^{n}log\left[\sum_{j=0}^{2}\varpi_{j|i}(p,q,D,\mathcal{M}_k)f_j(\mathbf{Y}_i|\mathbf{\Theta}_j,\mathbf{\Psi}_j)\right], \quad (4\text{–}7)$$

is given by

$$\bigtriangledown_{\Theta_j} \log L\left(\mathbf{\Theta}, \mathbf{\Psi}|\mathcal{M}, \mathbf{Y}\right) = \sum_{i=1}^{n} \frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2} \omega_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})} \times \bigtriangledown_{\Theta_j} log f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j),$$

$$\bigtriangledown_{\Psi} \log L(\mathbf{\Theta}, \mathbf{\Psi}|\mathcal{M}, \mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2} \varpi_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})} \times \bigtriangledown_{\Psi_j} \log f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j),$$

and

$$\bigtriangledown_{\varpi_{j|i}} \log L(\mathbf{\Theta}, \mathbf{\Psi}|\mathcal{M}, \mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2} \varpi_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})} \bigtriangledown_{\varpi_{j|i}} \log(\varpi_{j|i}).$$

Here we use $\bigtriangledown_{\varpi_{j|i}}$ to denote the gradient of log-likelihood with respect to $p,\ q$, and D.

Let

$$\mathbf{\Phi} = (p, q, D, \mathbf{\Theta}\prime, \mathbf{\Psi}\prime)\prime$$

Then to get the MLE of (4–7) we need to find the critical points that satisfy

$$\bigtriangledown_{\Phi} \log L(\mathbf{\Phi}|\mathcal{M}, \mathbf{Y}) = 0 \tag{4–8}$$

According to large sample theory we have the following existing theorem

Theorem 4.1. *(**Existence Theorem**) Suppose that f_j are non-degenerate normal distributions, and the Fisher information matrix $\mathbf{I}(\mathbf{\Phi})$ of (4–7) exists and is positive definite at the true parameter value $\mathbf{\Phi}_0$, then there exists a $\delta > 0$ such that for sufficiently large n, in the neighborhood $\{\mathbf{\Phi}|||\mathbf{\Phi} - \mathbf{\Phi}_0|| < \delta\}$ of $\mathbf{\Phi}_0$, with probability 1, there is a unique solution $\mathbf{\Phi}$ of (4–8) and $\sqrt{n}(\mathbf{\Phi} - \mathbf{\Phi}_0)$ has asymptotical distribution $N(\mathbf{0}, \mathbf{I}(\mathbf{\Phi}_0)^{-1})$.*

Proof It follows immediately from Redner and Walker [87] since our component distributions are all normally distributed.

To get the critical points the following Newton iteration is an attractive method

$$\mathbf{\Phi}_{t+1} = \mathbf{\Phi}_t - \alpha_k(\bigtriangledown^2 log L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}))^{-1} \bigtriangledown log L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}) \tag{4–9}$$

but it often gives us negative values for the estimates of the frequencies p, and q. At the same time the computation of Hessian causes a lot of mess. In order to avoid these embarrassing situations, we prefer to using EM algorithm to solve the problem. The general EM algorithm was originally proposed by Dempster, Laird, and Rubin [22] to solve the maximum-likelihood problems with missing data. The framework of using EM algorithm for solving maximum-likelihood estimates of finite mixture distributions can be found in [24, 66, 87] and their references. Here we briefly review the EM algorithm for our convenience. For the detail please see [22, 87], and its extensions [65].

Suppose our observed data are $\mathbf{Y}$, which are incomplete. And the complete data are $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, of which $\mathbf{Z}$ cannot be observed. Moreover, the MLE of $L(\mathbf{\Phi}|\mathbf{X})$ is much easier to solve than that of $L(\mathbf{\Phi}|\mathbf{Y})$. If we let $K(\mathbf{X}|\mathbf{Y}, \mathbf{\Phi})$ denote the conditional distribution of $\mathbf{X}$ given $(\mathbf{Y}, \mathbf{Z})$, and

$$Q(\mathbf{\Phi}|\mathbf{\Phi}_t) = E(logf(\mathbf{X}|\mathbf{\Phi})|\mathbf{Y}, \mathbf{\Phi}_t), \quad H(\mathbf{\Phi}|\mathbf{\Phi}_t) = E(logK(\mathbf{X}|\mathbf{Y}, \mathbf{\Phi})|\mathbf{Y}, \mathbf{\Phi}_t) \tag{4–10}$$

then the EM algorithm consists of the following two steps:

1. E-step. Compute $Q(\mathbf{\Phi}|\mathbf{\Phi}_t)$,
2. M-step. Compute $arg\ max\ Q(\mathbf{\Phi}|\mathbf{\Phi}_t)$.

Dempster, Laird, and Rubin (see [22]) have proved that $Q(\mathbf{\Phi}|\mathbf{\Phi}_t)$ is increasing, and $H(\mathbf{\Phi}|\mathbf{\Phi}_t)$ is decreasing. Wu (see [111]) has proved the convergence.

EM algorithm framework has been used to solve the MLE of the mixture models with distribution density function

$$f(\mathbf{y}_i; \mathbf{\Phi}) = \sum_{j=1}^{g} p_j f_j(\mathbf{y}_i; \mathbf{\Phi}_j) \tag{4–11}$$

by many authors(see [24, 65, 66, 87] and their references). In the rest of this paper we are going to show that the EM algorithm can be used to solve our problem (4–5), namely (4–1), and we also justify our choice of priors.

Let us define the variable

$$Z_{ikj} = 1 \text{ if object } i \text{ with marker } k \text{ belongs to genotype } j, \tag{4–12}$$

$$Z_{ikj} = 0 \text{ otherwise.} \tag{4–13}$$

Then Z_{ikj} for $k,\ j = 0,\ 1,\ 2$, and $\mathbf{Y}$ are our "complete" data. The corresponding likelihood function is

$$\begin{aligned} L_c(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathbf{Z}; \mathcal{M}) &= \prod_{k=0}^{2}\prod_{i=1}^{m_k}\prod_{j=0}^{2}\left[\omega_{j|k}(p, q, D, \mathcal{M}_k) f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j) P\left(\mathcal{M}_k|p\right)\right]^{Z_{ikj}} \\ &= \prod_{k=0}^{2}\prod_{i=1}^{m_k}\prod_{j=0}^{2}\left[\varpi_{j|k}(p, q, D, \mathcal{M}_k) f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j)\right]^{Z_{ikj}}, \end{aligned} \tag{4–14}$$

which gives us the log likelihood function

$$log L_c(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathbf{Z}; \mathcal{M}) = \sum_{k=0}^{2}\sum_{i=1}^{m_k}\sum_{j=0}^{2} Z_{ikj}\left[log\varpi_{j|k}(p, q, D) + log f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j)\right] \tag{4–15}$$

Now following the EM algorithm we have

$$Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(t)}) = \sum_{k=0}^{2}\sum_{i=1}^{m_k}\sum_{j=0}^{2} E(Z_{ikj})\left[log\varpi_{j|k}(p, q, D) + log f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j)\right], \tag{4–16}$$

and the next step is to maximize (4–16) with respect to our parameters. To do this we first find the $E(Z_{ikj})$ by using the following theorem.

Theorem 4.2. *In (4–16)*

$$E(Z_{ikj}) = \frac{\varpi_{j|k}(p_t, q_t, D_t) f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j^{(t)}, \boldsymbol{\Psi}_j^{(t)})}{\sum_{j\prime=0}^{2}\varpi_{j\prime|k}(p_t, q_t, D_t) f_{j\prime}(\mathbf{Y}_i|\boldsymbol{\Theta}_{j\prime}^{(t)}, \boldsymbol{\Psi}_{j\prime}^{(t)})}. \tag{4–17}$$

Proof For any single missing value Z_{ikj} with the observation $\mathbf{Y}_i$ the likelihood function is

$$L_c(\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j, p, q, D|\mathbf{Y}_i; Z_{ikj}; \mathcal{M}_k) = \left[\varpi_{j|k}(p, q, D, \mathcal{M}_k) f_j(\mathbf{Y}_i|\boldsymbol{\Theta}_j, \boldsymbol{\Psi}_j)\right]^{Z_{ikj}}, \tag{4–18}$$

and the corresponding log likelihood function is

$$logL_c(\mathbf{\Theta}_j, \mathbf{\Psi}_j, p, q, D|\mathbf{Y}_i; Z_{ikj}; \mathcal{M}_k) = Z_{ikj}\left[log\varpi_{j|k}(p, q, D, \mathcal{M}_k) + logf_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)\right] \quad (4\text{–}19)$$

Therefore,

$$E\left(\frac{\partial logL_c(\mathbf{\Theta}_j, \mathbf{\Psi}_j, p, q, D|\mathbf{Y}_i; Z_{ikj}; \mathcal{M}_k)}{\partial\varpi_{j|k}(p, q, D, \mathcal{M}_k)}\right) = E\left[Z_{ikj}\left(\frac{1}{\varpi_{j|k}(p, q, D, \mathcal{M}_k)}\right)\right] \quad (4\text{–}20)$$

On the other hand, if we do not include missing value Z_{ikj} for $\mathbf{Y}_i$ our likelihood function is

$$L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}_i; \mathcal{M}_k) = \sum_{j=0}^{2}\varpi_{j|k}(p, q, D, \mathcal{M}_k)f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j), \quad (4\text{–}21)$$

Hence,

$$E\left(\frac{\partial logL(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}_i; \mathcal{M}_k)}{\partial\varpi_{j|k}(p, q, D, \mathcal{M}_k)}\right) = E\left[\frac{\varpi_{j|k}f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2}\varpi_{j\prime|k}f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})}\left(\frac{1}{\varpi_{j|k}}\right)\right] \quad (4\text{–}22)$$

(4–20) and (4–22) are the two sides of the same coin. Following the definition of E-step we get (4–17).

Therefore, in the E step the posterior probability with which an individual i has a specific QTL genotype j based on the marker information $\mathcal{M}_k$ and phenotypic data is calculated by

$$\Omega_{j|i} = \frac{\varpi_{j|i}f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2}\varpi_{j\prime|i}f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})}. \quad (4\text{–}23)$$

The following theorem is very useful in our M-step computation.

Theorem 4.3. *In the M-step of EM algorithm the optimal solution of* p_{11}, p_{10}, p_{01} *and* p_{00} *are:*

$$\hat{p}_{11} = \frac{1}{2n}\left[\sum_{i=1}^{m_1}(2\Omega_{2|i} + \Omega_{1|i}) + \sum_{i=1}^{m_2}(\Omega_{2|i} + \xi\Omega_{1|i})\right] \quad (4\text{–}24)$$

$$\hat{p}_{10} = \frac{1}{2n}\left[\sum_{i=1}^{m_1}(\Omega_{1|i} + 2\Omega_{0|i}) + \sum_{i=1}^{m_2}\left(\Omega_{0|i} + (1-\xi)\Omega_{1|i}\right)\right] \quad (4\text{–}25)$$

$$\hat{p}_{01} = \frac{1}{2n}\left[\sum_{i=1}^{m_3}(2\Omega_{2|i} + \Omega_{1|i}) + \sum_{i=1}^{m_2}\left(\Omega_{2|i} + (1-\xi)\Omega_{1|i}\right)\right] \quad (4\text{–}26)$$

$$\hat{p}_{00} = \frac{1}{2n}\left[\sum_{i=1}^{m_3}(\Omega_{1|i} + 2\Omega_{0|i}) + \sum_{i=1}^{m_2}\left(\Omega_{0|i} + \xi\Omega_{1|i}\right)\right] \tag{4–27}$$

with

$$\xi = \frac{p_{11}p_{00}}{p_{11}p_{00} + p_{10}p_{01}}, \tag{4–28}$$

and the optimal solutions for p, q, *and* D *are:*

$$\hat{p} = \hat{p}_{11} + \hat{p}_{10}, \quad \hat{q} = \hat{p}_{11} + \hat{p}_{01}, \quad \hat{D} = \hat{p}_{11} - \hat{p}\hat{q} \tag{4–29}$$

Proof Note that (4–16) can be written as

$$\begin{aligned}
Q(\mathbf{\Phi}|\mathbf{\Phi}^{(t)}) &= \sum_{k=0}^{2}\sum_{i=1}^{m_k}\sum_{j=0}^{2} E(Z_{ikj}) log f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j) \\
&+ \sum_{i=1}^{m_1}\left\{E(Z_{i22})\left[log(p_{11}^2)\right] + E(Z_{i21})\left[log(2p_{11}p_{10})\right] + E(Z_{i20})\left[log(p_{10}^2)\right]\right\} \\
&+ \sum_{i=1}^{m_2}\left\{E(Z_{i12})\left[log(2p_{11}p_{01})\right] + E(Z_{i10})\left[log(2p_{10}p_{00})\right]\right\} \\
&+ \sum_{i=1}^{m_2}\left\{E(Z_{i11})\left[log(2p_{11}p_{00} + 2p_{10}p_{01})\right]\right\} \\
&+ \sum_{i=1}^{m_3}\left\{E(Z_{i02})\left[log(p_{01}^2)\right] + E(Z_{i01})\left[log(2p_{01}p_{00})\right] + E(Z_{i00})\left[log(p_{00}^2)\right]\right\}
\end{aligned}$$

Under the constrain condition

$$p_{11}^2 + 2p_{11}p_{10} + p_{10}^2 + 2p_{11}p_{01} + 2p_{11}p_{00} + 2p_{10}p_{01} + 2p_{10}p_{00} + p_{01}^2 + 2p_{01}p_{00} + p_{00}^2 = 1$$

the Lagranian function is

$$\begin{aligned}
\mathcal{L}(\mathbf{\Theta}, \mathbf{\Psi}, \mathbf{p}, \lambda) &= Q(\mathbf{\Phi}|\mathbf{\Phi}^{(t)}) + \lambda(p_{11}^2 + 2p_{11}p_{10} + p_{10}^2 + 2p_{11}p_{01} + 2p_{11}p_{00} \\
&+ 2p_{10}p_{01} + 2p_{10}p_{00} + p_{01}^2 + 2p_{01}p_{00} + p_{00}^2 - 1).
\end{aligned} \tag{4–30}$$

Differentiate (4–30) with respect p_{11}, p_{10}, p_{01}, and p_{00}, we get,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{11}} &= \sum_{i=1}^{m_1}\left\{E(Z_{i22})\frac{2}{p_{11}} + E(Z_{i21})\frac{1}{p_{11}}\right\} + \sum_{i=1}^{m_2} E(Z_{i12})\frac{1}{p_{11}} \\
&+ \sum_{i=1}^{m_2} E(Z_{i11})\frac{p_{00}}{p_{11}p_{00}+p_{10}p_{01}} + \lambda(2p_{11}+2p_{10}+p_{01}+2p_{00}) \\
&= \sum_{i=1}^{m_1}\left\{E(Z_{i22})\frac{2}{p_{11}} + E(Z_{i21})\frac{1}{p_{11}}\right\} + \sum_{i=1}^{m_2} E(Z_{i12})\frac{1}{p_{11}} \\
&+ \sum_{i=1}^{m_2} E(Z_{i11})\frac{p_{00}}{p_{11}p_{00}+p_{10}p_{01}} + 2\lambda, \qquad (4\text{–}31)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{10}} &= \sum_{i=1}^{m_1}\left\{E(Z_{i21})\frac{1}{p_{10}} + E(Z_{i20})\frac{2}{p_{10}}\right\} + \sum_{i=1}^{m_2} E(Z_{i10})\frac{1}{p_{10}} \\
&+ \sum_{i=1}^{m_2} E(Z_{i11})\frac{p_{01}}{p_{11}p_{00}+p_{10}p_{01}} + \lambda(2p_{11}+2p_{10}+2p_{01}+2p_{00}) \\
&= \sum_{i=1}^{m_1}\left\{E(Z_{i21})\frac{1}{p_{10}} + E(Z_{i20})\frac{2}{p_{10}}\right\} + \sum_{i=1}^{m_2} E(Z_{i10})\frac{1}{p_{10}} \\
&+ \sum_{i=1}^{m_2} E(Z_{i11})\frac{p_{01}}{p_{11}p_{00}+p_{10}p_{01}} + 2\lambda, \qquad (4\text{–}32)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{01}} &= \sum_{i=1}^{m_2}\left\{E(Z_{i12})\frac{1}{p_{01}} + E(Z_{i11})\frac{p_{10}}{p_{11}p_{00}+p_{10}p_{01}}\right\} \\
&+ \sum_{i=1}^{m_3}\left\{E(Z_{i02})\frac{2}{p_{01}} + E(Z_{i01})\frac{1}{p_{01}}\right\} + \lambda(2p_{11}+2p_{10}+2p_{01}+2p_{00}) \\
&= \sum_{i=1}^{m_2}\left\{E(Z_{i12})\frac{1}{p_{01}} + E(Z_{i11})\frac{p_{10}}{p_{11}p_{00}+p_{10}p_{01}}\right\} \\
&+ \sum_{i=1}^{m_3}\left\{E(Z_{i02})\frac{2}{p_{01}} + E(Z_{i01})\frac{1}{p_{01}}\right\} + 2\lambda, \qquad (4\text{–}33)
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p_{00}} &= \sum_{i=1}^{m_2}\left\{E(Z_{i10})\frac{1}{p_{00}} + E(Z_{i11})\frac{p_{11}}{p_{11}p_{00}+p_{10}p_{01}}\right\} \\ &+ \sum_{i=1}^{m_3}\left\{E(Z_{i01})\frac{1}{p_{00}} + E(Z_{i00})\frac{2}{p_{00}}\right\} + \lambda(2p_{11}+2p_{10}+2p_{01}+2p_{00}) \\ &= \sum_{i=1}^{m_2}\left\{E(Z_{i10})\frac{1}{p_{00}} + E(Z_{i11})\frac{p_{11}}{p_{11}p_{00}+p_{10}p_{01}}\right\} \\ &+ \sum_{i=1}^{m_3}\left\{E(Z_{i01})\frac{1}{p_{00}} + E(Z_{i00})\frac{2}{p_{00}}\right\} + 2\lambda. \end{aligned} \tag{4–34}$$

By KTT condition, we have

$$\hat{p}_{11} = \frac{1}{-2\lambda}\left[\sum_{i=1}^{m_1}\{2E(Z_{i22}) + E(Z_{i21})\} + \sum_{i=1}^{m_2}\left\{E(Z_{i12}) + E(Z_{i11})\frac{p_{11}p_{00}}{p_{11}p_{00}+p_{10}p_{01}}\right\}\right] \tag{4–35}$$

$$\hat{p}_{10} = \frac{1}{-2\lambda}\left[\sum_{i=1}^{m_1}\{E(Z_{i21}) + 2E(Z_{i20})\} + \sum_{i=1}^{m_2}\left\{E(Z_{i10}) + E(Z_{i11})\frac{p_{10}p_{01}}{p_{11}p_{00}+p_{10}p_{01}}\right\}\right] \tag{4–36}$$

$$\hat{p}_{01} = \frac{1}{-2\lambda}\left[\sum_{i=1}^{m_2}\left\{E(Z_{i12}) + E(Z_{i11})\frac{p_{10}p_{01}}{p_{11}p_{00}+p_{10}p_{01}}\right\} + \sum_{i=1}^{m_3}\{2E(Z_{i02}) + E(Z_{i01})\}\right] \tag{4–37}$$

and

$$\hat{p}_{00} = \frac{1}{-2\lambda}\left[\sum_{i=1}^{m_2}\left\{E(Z_{i10}) + E(Z_{i11})\frac{p_{11}p_{00}}{p_{11}p_{00}+p_{10}p_{01}}\right\} + \sum_{i=1}^{m_3}\{E(Z_{i01}) + 2E(Z_{i00})\}\right] \tag{4–38}$$

which complete the proof of (4–24), (4–25), (4–26), (4–27), and (4–28) since $\lambda = -n$.

The proof of (4–29) follows from the invariant theorem of maximum likelihood.

Note The formulas were obtained by Wang and Wu [106]. The proof has never been given as we know.

Also in the M step, the other parameters are estimated by solving the following log-likelihood equations:

$$\bigtriangledown_{\Theta_j} \log L\left(\mathbf{\Theta}, \mathbf{\Psi}|\mathcal{M}, \mathbf{Y}\right) = \sum_{i=1}^{n}\frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2}\varpi_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})} \cdot \bigtriangledown_{\Theta_j} log f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j) = 0 \tag{4–39}$$

$$\bigtriangledown_{\Psi} \log L(\mathbf{\Theta}, \mathbf{\Psi}|\mathcal{M}, \mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=0}^{2} \frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2} \varpi_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})} \cdot \bigtriangledown_{\Psi} \log f_j(\mathbf{Y}_i|\beth_j) = 0 \quad (4\text{–}40)$$

where $\beth_j = (\mathbf{\Theta}_j, \mathbf{\Psi}_j)$. Even for normal distributions it is often impossible for us to solve (4–39), and (4–40) explicitly. Therefore we have to turn to iterative method. Here we introduce the following iterative:

$$\mathbf{\Theta}_j^{(t+1)} = \mathbf{\Theta}_j^{(t)} - \alpha_t \left(\sum_{i=1}^{n} \Omega_{j|i} \bigtriangledown_{\Theta_j}^2 \log f_j(\mathbf{Y}_i|\mathbf{\Theta}_j^{(t)}, \mathbf{\Psi}_j^{(t)}) \right)^{-1} \bigtriangledown_{\Theta_j} \log L(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}_j^{(t)}|\mathbf{Y}) \quad (4\text{–}41)$$

$$\mathbf{\Psi}_j^{(t+1)} = \mathbf{\Psi}_j^{(t)} - \beta_t \left(\sum_{i=1}^{n} \sum_{j=0}^{2} \Omega_{j|i} \bigtriangledown_{\Psi_j}^2 \log f_j(\mathbf{Y}_i|\beth_j^{(t)}) \right)^{-1} \bigtriangledown_{\Psi_j} \log L(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}_j^{(t)}|\mathbf{Y}) \quad (4\text{–}42)$$

where α_t, and β_t are determined by the Armijo line search [3]. Of course, the method used in (4–41) and (4–42) is just conditional Newton Algorithm. A way to save computation is to use only the diagonal elements.

Let us summarize our algorithm as follows:

1. Use p_t, q_t, D_t, $\mathbf{\Theta}^{(t)}$, and $\mathbf{\Psi}^{(t)}$ to compute p_{t+1}, q_{t+1}, and D_{t+1} through (4–24-4–29),

2. Use p_{t+1}, q_{t+1}, D_{t+1}, $\mathbf{\Theta}^{(t)}$, and $\mathbf{\Psi}^{(t)}$ to compute $\mathbf{\Theta}_{t+1}$, and $\mathbf{\Psi}^{(t+1)}$ through (4–41), and (4–42),

We continue the above two steps until it converges.

Instead of using conditional Hessian matrices as we do in (4–41), and (4–42) we can hybrid Wang and Wu formulas with Newton Algorithm, and use

$$\begin{aligned} ((\mathbf{\Theta}^{(t+1)})\prime, \mathbf{\Psi}^{(t+1)})\prime &= ((\mathbf{\Theta}^{(t)})\prime, \mathbf{\Psi}^{(t)})\prime - \alpha_k (\bigtriangledown_{\Theta \cup \Psi}^2 logL(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}))^{-1} \\ &\cdot \bigtriangledown_{\Theta \cup \Psi} logL(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}) \end{aligned} \quad (4\text{–}43)$$

4.2.4 Main Results

Our modified algorithm is actually a Rao-Blackerzation. We have the following theorem

Theorem 4.4. *If (4–24-4–29) are used to compute p, q, and D, then the sampling errors by using (4–41-4–42) to solve the critical points of (4–7) are smaller than that of Newton method (4–43) if both of them are convergent*

Proof Let us define binary latent variables Z_{ij} as

$$P(Z_{ij} = 1) = \Omega_{j|i} = \frac{\varpi_{j|i} f_j(\mathbf{Y}_i|\mathbf{\Theta}_j, \mathbf{\Psi}_j)}{\sum_{j\prime=0}^{2} \varpi_{j\prime|i} f_{j\prime}(\mathbf{Y}_i|\mathbf{\Theta}_{j\prime}, \mathbf{\Psi}_{j\prime})}$$

Treat solving the critical points of (4–7) as a random process Y. Then

$$Var(Y) = Var(E(Y|\mathbf{Z})) + E(Var(Y|\mathbf{Z})) \tag{4–44}$$

where $\mathbf{Z}$ denotes the collection of all Z_{ij} for $i = 1, \ldots, n$, and $j = 0, 1, 2$. The Hessian of (4–41-4–42) is conditional on $\mathbf{Z}$. Therefore its variance is $Var(E(Y|\mathbf{Z}))$ but that of (4–43) is $Var(E(Y|\mathbf{Z})) + E(Var(Y|\mathbf{Z}))$.

Then, we prove the estimators of p, q, and D are consistent. We see it is a trade off between Newton algorithm and EM. EM is slower than Newton but its error is smaller. Second, we prove the estimators of p, q, and D are consistent.

Theorem 4.5. *If $\hat{p}$, $\hat{q}$, and $\hat{D}$ are given by (4–29), then $\hat{p} \to p$, $\hat{q} \to q$, and $\hat{D} \to D$ as $n \to \infty$. Also, $\hat{p}$ is equal to the MLE of multinomial distribution $\frac{n!}{m_1!m_2!m_3!}(p^2)^{m_1}(2p(1-p))^{m_2}((1-p)^2)^{m_3}$ with observation (m_1, m_2, m_3). Furthermore, if $\hat{p}_{11}$, $\hat{p}_{10}$, $\hat{p}_{01}$, and $\hat{p}_{00}$ are defined by (4–24), (4–25), (4–26), and (4–27), then $\hat{p}_{11} \to pq + D$, $\hat{p}_{10} \to p(1-q) - D$, $\hat{p}_{01} \to (1-p)q - D$, and $\hat{p}_{00} \to (1-p)(1-q) + D$, as $n \to \infty$.*

Proof According to the definition of $\Omega_{2|i}$, we know

$$\sum_{i=1}^{m_k} (\Omega_{2|i} + \Omega_{1|i} + \Omega_{0|i}) = 1 \tag{4–45}$$

for $k = 0, 1, 2$. We also know

$$lim_{n\to\infty} \frac{m_1}{n} = p^2 \tag{4–46}$$

$$lim_{n\to\infty} \frac{m_2}{n} = 2p(1-p) \tag{4–47}$$

Therefore, using (4–24), and (4–25), we have

$$\begin{aligned}
\hat{p} &= \hat{p}_{11} + \hat{p}_{10} \\
&= \frac{1}{2n}\left[\sum_{i=1}^{m_1}(2\Omega_{2|i} + \Omega_{1|i}) + \sum_{i=1}^{m_2}(\Omega_{2|i} + \xi\Omega_{1|i})\right] \\
&+ \frac{1}{2n}\left[\sum_{i=1}^{m_1}(\Omega_{1|i} + 2\Omega_{0|i}) + \sum_{i=1}^{m_2}\left(\Omega_{0|i} + (1-\xi)\Omega_{1|i}\right)\right] \\
&= \frac{1}{2n}\left[2\sum_{i=1}^{m_1}(\Omega_{2|i} + \Omega_{1|i} + \Omega_{0|i}) + \sum_{i=1}^{m_2}(\Omega_{2|i} + \Omega_{1|i} + \Omega_{0|i})\right] \\
&= \frac{1}{2n}\left[2m_1 + m_2\right] \\
&\to p^2 + p(1-p) \quad (\text{as} \quad n \to \infty) \\
&= p
\end{aligned}$$

which completes the proof of $\hat{p} \to p$, and also shows that $\hat{p} = \frac{2m_1+m_2}{2n}$, namely the MLE of the multinomial distribution $MN\left(n; p^2, 2p(1-p), (1-p)^2\right)$.

To prove $\hat{q} \to q$ let us assume the genotype numbers of AA, Aa, and aa are n_1, n_2, and n_3, respectively, then the likelihood of joint distribution can be written as

$$\begin{aligned}
L(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathcal{A}) &= \prod_{i=1}^{n_1}\left(q^2 f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2)\right)\prod_{i=1}^{n_2}(2q(1-q)f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2)) \\
&\cdot \prod_{i=1}^{n_3}\left((1-q)^2 f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2)\right) \qquad (4\text{–}48)
\end{aligned}$$

Then the log likelihood function becomes

$$\begin{aligned}
log L(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathcal{A}) &= n_1 log(q^2) + \sum_{i=1}^{n_1} log f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2) \\
&+ n_2 log(2q(1-q)) + \sum_{i=1}^{n_2} log f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1) \\
&+ n_3 log((1-q)^2) + \sum_{i=1}^{n_3} log f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0)
\end{aligned}$$

with

$$n_1 = \sum_{i=1}^{n}\sum_{k=0}^{2} Z_{ik2}, \quad n_2 = \sum_{i=1}^{n}\sum_{k=0}^{2} Z_{ik1}, \quad n_3 = \sum_{i=1}^{n}\sum_{k=0}^{2} Z_{ik0}. \tag{4–49}$$

By **Theorem4.2**,

$$n_1 = \sum_{i=1}^{n} E(\Omega_{2|i}). \tag{4–50}$$

So we can use

$$\sum_{i=1}^{n} \Omega_{2|i} = \sum_{i=1}^{m_1} \Omega_{2|i} + \sum_{i=1}^{m_2} \Omega_{2|i} + \sum_{i=1}^{m_3} \Omega_{2|i} \tag{4–51}$$

to estimate n_1.

Similarly, we can use

$$\sum_{i=1}^{n} \Omega_{1|i} = \sum_{i=1}^{m_1} \Omega_{1|i} + \sum_{i=1}^{m_2} \Omega_{1|i} + \sum_{i=1}^{m_3} \Omega_{1|i} \tag{4–52}$$

to estimate n_2. On the other hand, according to (4–24), (4–26), and (4–29) we have

$$\begin{aligned}
\hat{q} &= p_{11} + p_{01} \\
&= \frac{1}{2n}\left[\sum_{i=1}^{m_1}(2\Omega_{2|i} + \Omega_{1|i}) + \sum_{i=1}^{m_2}(\Omega_{2|i} + \xi\Omega_{1|i})\right] \\
&+ \frac{1}{2n}\left[\sum_{i=1}^{m_3}(2\Omega_{2|i} + \Omega_{1|i}) + \sum_{i=1}^{m_2}\left(\Omega_{2|i} + (1-\xi)\Omega_{1|i}\right)\right] \\
&= \frac{1}{2n}\left[2\left(\sum_{i=1}^{m_1}\Omega_{2|i} + \sum_{i=1}^{m_2}\Omega_{2|i} + \sum_{i=1}^{m_3}\Omega_{2|i}\right)\right] \\
&+ \frac{1}{2n}\left[\sum_{i=1}^{m_1}\Omega_{1|i} + \sum_{i=1}^{m_2}\Omega_{1|i} + \sum_{i=1}^{m_3}\Omega_{1|i}\right],
\end{aligned}$$

which can be used to estimate

$$\frac{1}{2n}(2n_1) + \frac{1}{2n}(n_2) \longrightarrow^{P} q^2 + q(1-q) = q \ \ (as \ \ n \longrightarrow \infty). \tag{4–53}$$

To prove $\hat{D} \rightarrow^{P} D \ \ (as \ \ n \rightarrow \infty)$ we assume that the genotype numbers of AA, Aa, and aa are m_{11}, m_{12}, and m_{13}, respectively, among the marker type MM, then

$$m_1 = m_{11} + m_{12} + m_{13}. \tag{4–54}$$

Therefore, the likelihood function for the marker type *MM* ONLY should be

$$\begin{aligned}
L(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}_2) &= \prod_{i=1}^{m_{11}} \left[\frac{(pq+D)^2}{p^2} f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) P(\mathcal{M}_2) \right] \\
&\times \prod_{i=1}^{m_{12}} \left[\frac{2(pq+D)(p(1-q)-D)}{p^2} f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) P(\mathcal{M}_2) \right] \\
&\times \prod_{i=1}^{m_{13}} \left[\frac{(p(1-q)-D)^2}{p^2} f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0) P(\mathcal{M}_2) \right] \\
&= \prod_{i=1}^{m_{11}} \left[(pq+D)^2 f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) \right] \\
&\times \prod_{i=1}^{m_{12}} \left[2(pq+D)(p(1-q)-D) f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) \right] \\
&\times \prod_{i=1}^{m_{13}} \left[(p(1-q)-D)^2 f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0) \right]
\end{aligned}$$

whose log likelihood function is

$$\begin{aligned}
&2m_{11} log(pq+D) + \sum_{i=1}^{m_{11}} log f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) + m_{12}[log(pq+D) + log(p(1-q)-D)] \\
&+ \sum_{i=1}^{m_{12}} log f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) + 2m_{13} log(p(1-q)-D) + \sum_{i=1}^{m_{13}} f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0). \quad (4\text{–}55)
\end{aligned}$$

which means

$$\begin{aligned}
m_{11} &= \textstyle\sum_{i=1}^{m_1} Z_{i22} = \sum_{i=1}^{m_1} E(\Omega_{2|i}), \\
m_{12} &= \textstyle\sum_{i=1}^{m_1} Z_{i21} = \sum_{i=1}^{m_1} E(\Omega_{1|i}), \\
m_{13} &= \textstyle\sum_{i=1}^{m_1} Z_{i20} = \sum_{i=1}^{m_1} E(\Omega_{0|i}).
\end{aligned} \quad (4\text{–}56)$$

If we let m_{21}, m_{22}, m_{23}, and m_{24} denote the genotype numbers of AA, Aa, aA, and aa among the marker type Mm, then consider

$$\begin{aligned}
L(\boldsymbol{\Theta}, \boldsymbol{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M}_1) &= \prod_{i=1}^{m_{21}} \left[\frac{2(pq+D)((1-p)q-D)}{2p(1-p)} f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) P(\mathcal{M}_1) \right] \\
&\times \prod_{i=1}^{m_{22}} \left[\frac{2(pq+D)((1-p)(1-q)+D)}{2p(1-p)} f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) P(\mathcal{M}_1) \right] \\
&\times \prod_{i=1}^{m_{23}} \left[\frac{2(p(1-q)-D)((1-p)q-D)}{2p(1-p)} f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) P(\mathcal{M}_1) \right] \\
&\times \prod_{i=1}^{m_{24}} \left[\frac{(p(1-q)-D)((1-p)(1-q)+D)}{2p(1-p)} f_0(\mathbf{Y}_i) P(\mathcal{M}_1) \right] \\
&= \prod_{i=1}^{m_{21}} [2(pq+D)((1-p)q-D) f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2)] \\
&\times \prod_{i=1}^{m_{22}} [2(pq+D)((1-p)(1-q)+D) f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1)] \\
&\times \prod_{i=1}^{m_{23}} [2(p(1-q)-D)((1-p)q-D) f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1)] \\
&\times \prod_{i=1}^{m_{24}} [(p(1-q)-D)((1-p)(1-q)+D) f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0)]
\end{aligned}$$

with log likelihood function

$$\begin{aligned}
& m_{21} log[2(pq+D)((1-p)q-D)] + \sum_{i=1}^{m_{21}} log f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) + m_{22} log[2(pq+D) \\
& \cdot((1-p)q-D)] + \sum_{i=1}^{m_{22}} log f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) + m_{23} log[2(p(1-q)-D)((1-p)q-D)] \\
& + \sum_{i=1}^{m_{23}} f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) + m_{24} log[2(p(1-q)-D)((1-p)(1-q)+D)] + \sum_{i=1}^{m_{24}} f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0) \\
&= \sum_{i=1}^{m_2} [2(pq+D)((1-p)q-D) f_2(\mathbf{Y}_i|\boldsymbol{\Theta}_2, \boldsymbol{\Psi}_2) + (2(pq+D)((1-p)(1-q)+D) \\
&\quad +2(p(1-q)-D) \cdot ((1-p)q-D)) f_1(\mathbf{Y}_i|\boldsymbol{\Theta}_1, \boldsymbol{\Psi}_1) \\
&\quad +2(p(1-q)-D)((1-p)(1-q)+D) f_0(\mathbf{Y}_i|\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0)].
\end{aligned}$$

Hence, we get

$$m_{21} = \sum_{i=1}^{m_2} Z_{i12} = \sum_{i=1}^{m_2} E(\Omega_{2|i}), \tag{4–57}$$

$$m_{22} + m_{23} = \sum_{i=1}^{m_2} Z_{i11} = \sum_{i=1}^{m_2} E(\Omega_{1|i}), \tag{4–58}$$

and

$$m_{24} = \sum_{i=1}^{m_2} Z_{i10} = \sum_{i=1}^{m_2} E(\Omega_{0|i}). \tag{4–59}$$

Similarly, if we assume that m_{31}, m_{32}, and m_{33} represent the genotype numbers of AA, Aa, and aa among the marker mm, we have

$$\begin{aligned} m_{31} &= \textstyle\sum_{i=1}^{m_3} Z_{i02} = \sum_{i=1}^{m_3} E(\Omega_{2|i}), \\ m_{32} &= \textstyle\sum_{i=1}^{m_3} Z_{i01} = \sum_{i=1}^{m_3} E(\Omega_{1|i}), \\ m_{33} &= \textstyle\sum_{i=1}^{m_3} Z_{i00} = \sum_{i=1}^{m_3} E(\Omega_{0|i}). \end{aligned} \tag{4–60}$$

For multinomial distribution $MN(n;\ \pi_1, \ldots, \pi_{10})$ with

$$\pi_1 = (pq + D)^2, \tag{4–61}$$

$$\pi_2 = 2(pq + D)(p(1-q) - D), \tag{4–62}$$

$$\pi_3 = (p(1-q) - D)^2, \tag{4–63}$$

$$\pi_4 = 2(pq + D)((1-p)q - D), \tag{4–64}$$

$$\pi_5 = 2(pq + D)((1-p)(1-q) + D), \tag{4–65}$$

$$\pi_6 = 2(p(1-q) - D)((1-p)q - D), \tag{4–66}$$

$$\pi_7 = 2(p(1-q) - D)((1-p)(1-q) + D), \tag{4–67}$$

$$\pi_8 = ((1-p)q - D)^2, \tag{4–68}$$

$$\pi_9 = 2((1-p)q - D)((1-p)(1-q) + D), \tag{4–69}$$

$$\pi_{10} = ((1-p)(1-q) + D)^2. \tag{4–70}$$

and observation $(m_{11}, m_{12}, m_{13}, m_{21}, m_{22}, m_{23}, m_{24}, m_{31}, m_{32}, m_{33}, m_{41}, m_{42}, m_{43})$, by invariant theorem the *MLE* for $pq + D$ is

$$\hat{p}_{11} = \widehat{pq + D} = \frac{2m_{11} + m_{12} + m_{21} + m_{22}}{2n}, \tag{4–71}$$

since

$$(pq + D)^2 + (pq + D)(p(1-q) - D) + (pq + D)((1-p)q - D) + (pq + D) \cdot((1-p)(1-q) + D) = pq + D.$$

Similarly, we have

$$\hat{p}_{10} = p\widehat{(1-q)} - D = \frac{m_{12} + 2m_{13} + m_{23} + m_{24}}{2n}, \tag{4–72}$$

due to

$$(pq + D)(p(1-q) - D) + (p(1-q) - D)^2 + (p(1-q) - D)((1-p)q - D) + (p(1-q) - D)((1-p)(1-q) + D) = p(1-q) - D,$$

and

$$\hat{p}_{01} = \widehat{(1-p)q} - D = \frac{m_{21} + m_{23} + 2m_{31} + m_{32}}{2n}, \tag{4–73}$$

since

$$(pq+D)((1-p)q-D)+(p(1-q)-D)((1-p)q-D)+((1-p)q-D)^2+$$
$$((1-p)q-D)((1-p)(1-q)+D)=(1-p)q-D,$$

and

$$\hat{p}_{00}=\widehat{(1-p)(1-q)}+D=\frac{m_{22}+m_{24}+m_{32}+2m_{30}}{2n}, \tag{4–74}$$

because

$$(pq+D)((1-p)(1-q)+D)+(p(1-q)-D)((1-p)(1-q)+D)+((1-p)q-D)$$
$$\cdot((1-p)(1-q)+D)+((1-p)(1-q)+D)^2=(1-p)(1-q)+D.$$

Therefore

$$\hat{p}_{11}\longrightarrow^{P} pq+D \quad (as \;\; n\longrightarrow\infty) \tag{4–75}$$

$$\hat{p}_{01}\longrightarrow^{P} (1-p)q-D \quad (as \;\; n\longrightarrow\infty) \tag{4–76}$$

$$\hat{p}_{10}\longrightarrow^{P} p(1-q)-D \quad (as \;\; n\longrightarrow\infty) \tag{4–77}$$

$$\hat{p}_{00}\longrightarrow^{P} (1-p)(1-q)+D \quad (as \;\; n\longrightarrow\infty). \tag{4–78}$$

Hence

$$\hat{D}=\hat{p}_{11}-\hat{p}\hat{q}\longrightarrow^{P} D \quad (as \;\; n\longrightarrow\infty). \tag{4–79}$$

But the problem is we can not estimate m_{22}, and m_{23} directly since we can only estimate their sum. Therefore, in the iterative algorithm (4–28) is defined, which involves the

second layer of the EM algorithm for "complete" data likelihood

$$\prod_{i=1}^{m_{22}} (2(pq + D)((1-p)(1-q) + D))\, f_1(\mathbf{Y}_i|\mathbf{\Theta}_1, \mathbf{\Psi}_1)$$
$$\cdot \prod_{i=1}^{m_{23}} (2(p(1-q) - D)((1-p)q - D))\, f_1(\mathbf{Y}_i|\mathbf{\Theta}_1, \mathbf{\Psi}_1)$$

and "incomplete" data likelihood

$$\prod_{i=1}^{m_{22}+m_{23}} (2(pq + D)((1-p)(1-q) + D) + 2(p(1-q) - D)((1-p)q - D))\, f_1(\mathbf{Y}_i|\mathbf{\Theta}_1, \mathbf{\Psi}_1)$$

respectively. The details have been omitted.

Well, for univariate distribution, or the variance and covariance matrices have nice structure, we do not always recommend to use (4–42) to solve for $\mathbf{\Psi}$. Since in univariate case we can easily get the unbiased estimators of variance parameters.

Theorem 4.6. *(**Capture Theorem**) Let f_j be non-degenerate normal distributions and Ω be a closed bounded parameter subspace containing the MLE*

$$(p^*, q*, D^*, \mathbf{\Theta}^{(*0)}, \mathbf{\Theta}^{(*1)}, \mathbf{\Theta}^{(*2)}, \mathbf{\Psi}^*)$$

of $L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M})$. If $(p^*, q*, D^*, \mathbf{\Theta}^{(*0)}, \mathbf{\Theta}^{(*1)}, \mathbf{\Theta}^{(*2)}, \mathbf{\Psi}^*)$ is the only critical point of $L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M})$ in Ω and the sequence $\{(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}^{(t)}, p_t, q_t, D_t)\}$ generated by the above algorithm always in Ω for all t. Then $(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}^{(t)}, p_t, q_t, D_t) \to (\mathbf{\Theta}^*, \mathbf{\Psi}^*, p, q^*, D^*)$ as $t \to \infty$. Furthermore, $||(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}^{(t)}, p_t, q_t, D_t) - (\mathbf{\Theta}^*, \mathbf{\Psi}^*, p, q^*, D^*)||$ converges to zero linearly.

Proof Since $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(t)})$ is continuous with respect to both $\mathbf{\Phi}$ and $\mathbf{\Phi}^{(t)}$,

$$\{(\mathbf{\Theta}^{(t)}, \mathbf{\Psi}^{(t)}, p_t, q_t, D_t)\}$$

converges to a stationary point of $L(\mathbf{\Theta}, \mathbf{\Psi}, p, q, D|\mathbf{Y}; \mathcal{M})$, which must be

$$(p^*, q*, D^*, \mathbf{\Theta}^{(*0)}, \mathbf{\Theta}^{(*1)}, \mathbf{\Theta}^{(*2)}, \mathbf{\Psi}^*)$$

in Ω. The second part follows [69, 70].

4.3 DSE Data Analysis

DSE is a data set about the genetic study of body mass index (BMI) for different genotypes. We are going to study its relationship with different markers. The density functions for the components of the mixture model are

$$f_j(y_i|\mu_j,\sigma^2)=\frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(y_i-\mu_j)^2}{2\sigma^2}\right\},\quad j=1,\ 2,\ 3, \tag{4–80}$$

Therefore in the M-step of EM algorithm we have

Theorem 4.7. *Under the assumption of (4–80)*

$$\hat{\mu_j}=\frac{\sum_{k=0}^{2}\sum_{i=1}^{m_k}\Omega_{j|ik}y_i}{\sum_{k=0}^{2}\sum_{i=1}^{m_k}\Omega_{j|ik}},\quad \hat{\sigma^2}=\frac{\sum_{k=0}^{2}\sum_{i=1}^{m_k}\sum_{j=0}^{2}\Omega_{j|ik}(y_i-\hat{\mu_j})^2}{n}. \tag{4–81}$$

Proof Using (4–16) and (4–23) we have

$$\frac{\partial\mathcal{L}}{\partial\mu_j}=\frac{\partial Q(\mu,\sigma^2|\mu^{(t)},\sigma^{2(t)})}{\partial\mu_j}=\sum_{k=0}^{2}\sum_{i=1}^{m_k}\Omega_{j|ik}\left(\frac{y_i-\mu_j}{\sigma^2}\right), \tag{4–82}$$

and

$$\frac{\partial\mathcal{L}}{\partial(\sigma^2)}=\frac{\partial Q(\mu,\sigma^2|\mu^{(t)},\sigma^{2(t)})}{\partial(\sigma^2)}=\sum_{k=0}^{2}\sum_{i=1}^{m_k}\sum_{j=0}^{2}\Omega_{j|ik}\left(-\frac{1}{2\sigma^2}+\frac{(y_i-\mu_j)^2}{2(\sigma^2)^2}\right). \tag{4–83}$$

$\frac{\partial\mathcal{L}}{\partial\mu}=0$ and $\frac{\partial\mathcal{L}}{\partial(\sigma^2)}=0$ give us (4–81).

Table 4-1. Estimates of the Parameters for five codons in DSE data

Parameters	Codon16	Codon27	Codon49	Codon398	Codon492
p	0.6059	0.6154	0.8429	0.7384	0.5581
q	0.9469	0.9294	0.9542	0.9667	0.9165
D	0.0173	-0.0272	0.0131	-0.0257	0.0466
μ_1	28.2374	27.8987	28.2138	28.6255	29.4148
μ_2	41.3258	40.1542	42.4086	44.7031	29.1139
μ_3	100.8900	100.8900	100.8900	100.8900	100.8900
σ^2	30.1626	28.2537	44.2228	31.2721	59.184

4.4 Numerical Experiment

After proving the above theorems we must do numerical experiment to check our results. We simulate 200 times at the sample size 200 for heritabilities of $H^2 = 0.1$ and $H^2 = 0.4$. Simulation is consistent with our theorems. We follow the usual definition for computation error:

$$ComputationError = \frac{||TrueValue - Estimate||}{TrueValue}$$

although this may cause problem when true value is zero. **Table 4.4** shows that **Theorem 4.4** is correct. But we must note here that **Theorem 4.4** holds only if both methods follow the same computation procedures, namely, the optimization methods must be the same. If one is gradient method and the other is not, then there is no reason that **Theorem 4.4** will hold since the computation process is not the same.

Table 4-2. The MLEs of parameters for three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.1$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Phenotypic Parameters						
	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
μ	21	21.3902(0.5057)	25	22.8251(0.6833)	31	31.2811(3.5474)

Genetic Parameters and Variances		
	Given	MLE
p	0.56	0.5604(0.0406)
q	0.89	0.8597(0.0680)
D	0.03	0.0460(0.0108)
σ^2	31.36	32.1692(0.0104)

Table 4-3. The MLEs of parameters for three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.4$. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

Phenotypic Parameters						
	AA		*Aa*		*aa*	
	Given	MLE	Given	MLE	Given	MLE
μ	21	21.0125(0.2373)	25	23.9699(0.4672)	31	29.9134(0.5133)

Genetic Parameters and Variances		
	Given	MLE
p	0.56	0.5586(0.0226)
q	0.89	0.8558(0.0370)
D	0.03	0.0382(0.0173)
σ^2	4.9723	5.334(0.5133)

Table 4-4. Comparison the MLEs of Newton method and EM algorithm for Estimating the parameters of three different QTL genotypes, and the association between the marker and QTL in a natural population assuming that the heritability of the assumed QTL is $H^2 = 0.1$. The number in the parentheses are the computation errors.

Parameters	True Value	Newton	EM
p	0.56	0.6681(0.1930)	0.5605(0.0008)
q	0.89	0.7114(0.2007)	0.8619(0.0316)
D	0.03	0.0452(0.5067)	0.0363(0.2100)
μ_1	21	24.9405(0.1876)	21.3798(0.0181)
μ_2	25	19.9910(0.2004)	22.8305(0.0868)
μ_3	31	37.8412(0.2207)	31.3667(0.0118)
σ^2	31.36	34.6325(0.1044)	31.7875(0.0136)

4.5 Conclusion and Discussion

In this chapter we focus on the solution of mixture models used in solving complex genetic traits in the past years. We first prove the existence theorem, which is a simple corollary of [87]. Then we strictly follow EM algorithm and provide the E-step in our second theorem. In **Theorem 4.3** we give the solution for genetic parameters p, q, and D. In the past years people always say that EM algorithm is slow. Here we show it will gives us smaller computation errors (**Theorem 4.4**). **Theorem 4.5** or consistent theorem show that the accuracy of p, q, and D are driven by sample size and the accuracy of other parameters. More research results and computations tips will be given in the follow up research.

CHAPTER 5
ONGOING RESEARCH AND FUTURE WORK

Besides writing this dissertation I also have coauthored several published papers with Dr. Hager, and Dr. Wu. Another several papers have been accepted to be published. Three papers are in the process of being completed. In the past three years I have given five posters, papers, or talks at various conferences and a total of several thousand dollars have been awarded to me to support my travels. Right now we are focusing on the following projects:

1. Family Based Linkage Disequilibrium: We are going to present a statistical algorithm for constructing a joint linkage-linkage disequilibrium map by simultaneously estimating the recombination fractions and linkage disequilibria using multilocus marker data in a natural human population. The data are a set of random unrelated families, each including a father, a mother and a varying number of offspring, sampled from a population at Hardy-Weinberg equilibrium. The strategy is to provide an algorithm and study its theoretical and practical properties.

2. Sequencing Complex Diseases: We are going to study specific sequence variants that are responsible for disease risk based on the haplotype structure provided by HapMap. As an example we are going to model the data from a human obesity study with 155 patients.

3. Soybean data–Application of our methodology: As an application of Chapter 1 we are going to study Soybean Data and model its biological development.

    (a) The data were collected by a group of Chinese scientists. Sample size is 184. The data contain the following information:

        i. 25 linkage segments,

        ii. 498 markers (genes) for each sample,

        iii. Distances between markers.

(b) Soybean follows the following growth dynamic model:

$$\begin{aligned}
\frac{dM_L}{dt} &= \alpha_L M_T^{\beta_L} - \lambda_L M_L, \\
\frac{dM_P}{dt} &= \alpha_P M_T^{\beta_P}, \\
\frac{dM_S}{dt} &= \alpha_S M_T^{\beta_S}, \\
\frac{dM_R}{dt} &= \alpha_R M_T^{R} - \lambda_R M_R.
\end{aligned} \tag{5–1}$$

where M_L, M_P, M_S, and are the biomass of leaves, petioles, and stems,respectively; and M_T is the total biomass. All the Ms were collected at different time points. Our research goal for this project is to use our method and interval functional mapping and find the genes affecting the soybean growth dynamic model (5–1).

4. Decay dynamics of HIV-1: We plant to use use Markov chain model to characterize specific genes or quantitative trait loci (QTLs) that affect viral life trajectories within the framework of a dynamic system. The purpose is to focus on different genes for different stage.

In the future we are going to study:

1. Bayesian method: Use Bayesian approach to study linkage and put prior on our p, and q.

2. RNA related problem.

3. Protein.

REFERENCES

[1] Atchley WR, Xu S, Cowley DE(1997) "Altering developmental trajectories in mice by restricted index selection", *Genetics* 146: 629-640.

[2] Aylor DL, Zeng Z-B(2008) "From classic genetics to quantitative genetics to systems biology: Modeling epistasis", *PLoS Genet.* 4(3).

[3] Bertsekas DP(2003) "Nonlinear Programming," Athena Scientific, 2nd, Belmont, MA.

[4] Bonhoeffer S(1998) "Models of viral kinetics and drug resistance in HIV-1 infection," *AIDS Patient Care Stds.* , vol. 12, pp.769–774.

[5] Bonhoeffer S, Coffin JM, Nowak MA(1997) "Human immuodeficiency virus drug therapy and virus load," *J. Virology*, vol. 71, pp.6971–6976.

[6] Bonhoeffer S, May RM, Shaw GM, Nowak MA(1999) "Virus dynamics and drug therapy," *Proc. Natl. Acad. Sci. USA*, vol. 94, pp.6971–6976.

[7] Bonhoeffer S, Nowak MA(1997) "Pre-existence and emergence of drug resistance in HIV-1 infection," *Proc. Roy. Soc. Lond B*, vol. 264, pp.631–637.

[8] Bremaud P(1999), *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer, New York.

[9] Broman KW, Speed TP (2002) " A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion)", *J. Roy. Stat. Soc. B*, 64: 641-656.

[10] Carlborg O, Haley CS(2004) "Epistasis: too often neglected in complex trait studies?" *Nat. Rev. Genet.* 5: 618-625.

[11] Chang MM, Wu RL, Wu S, Casella G(2009), "Score statistics of quantitative trait locus mapping", *Stat. Appl. Genet. Mol. Biol.* 8: Iss. 1, Article 16.

[12] Chen J, Reynolds J(1997) " A coordination model of carbon allocation in relation to water supply", *Ann. Bot.* 80: 45-55.

[13] Chen J, Wu H(2008) "Efficient Local Estimation for Time-varying Coefficients in Deterministic Dynamic Models with Applications to HIV-1 Dynamics", *J. Am. Stat. Assoc.* 103: 369-384.

[14] Chen J, Wu H(2008) " Estimation of time-varying parameters in deterministic dynamic models with application to HIV infections", *Stat. Sinica* 18: 987-1006.

[15] Churchill GA, Doerge RW(1994), "Empirical threshold values for quantitative trait mapping," *Genetics*, vol. 138, pp.963–971.

[16] Clark AG(1990) “Inference of haplotypes from PCR-amplified samples of diploid populations”, *Mol. Biol. Evol.* 7: 111-122.

[17] Clayton TA, Lindon JC, Cloarec O, Antti H, Charuel C et al.(2006) “ Pharmaco-metabonomic phenotyping and personalized drug treatment”, *Nature* 440: 1073-1077.

[18] Cockerham CC(1954) “An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present”, *Genetics* 39: 859-882.

[19] Cooper RS, Psaty BM(2003) “ Genomics and medicine: Distraction, incremental progress, or the dawn of a new age?” *Ann. Internal Med.*. 138: 576-680.

[20] Cui YH, Zhu J, Wu RL(2006) “ Functional mapping for genetic control of programmed cell death”, *Physiol. Genom* 25: 458-469.

[21] Deeks SG(2003) ”Treatment of antiretroviral-drug-resistant HIV-1 infection”, *Lancet*, 362: 2002–2011.

[22] Dempster AP, Laird NM, Rubin DB(1977) “Maximum likelihood from incomplete data via EM algorithm”, *J. Roy. Stat. Soc. B*, vol. 39, pp.1–38.

[23] Diggle PJ, Verbyla AP(1998) “ Nonparametric estimation of covariance structure in longitudinal data”, *Biometrics* 54: 401-415.

[24] Everitt BS, Hand DJ(1981) “Finite Mixture Distributions”, Chapman and Hall, New York.

[25] Fan J, Huang T, Li RZ(2007) “Analysis of longitudinal data with semiparametric estimation of covariance function”, *J. Am. Stat. Assoc.* 35: 632-641.

[26] Fan J, Li R(2004) “ New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis”, *J. Am. Stat. Assoc.* 99:710-723.

[27] Fan J, Zhang J(2000) “Two-step estimation of functional linear models with applications to longitudinal data”, *J. Roy. Stat. Soc. B* 62: 303-322.

[28] Feng P, Navaratna M(2007), “Modelling periodic oscillations during somitogenesis”, *Math. Biosci. Engin.* 4: 661-673.

[29] Frank SA(2004) “Age-specific acceleration of cancer”, *Curr. Biol.* 14: 242-246.

[30] Gelman CP, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis*, Chapman & Hall, London.

[31] Glasbey CA (1988) “ Standard errors resilient to error variance misspecification”, *Biometrika* 75: 201-206.

[32] Goldbeter A(2002) " Computational approaches to cellular rhythms", *Nature* 420: 238-245.

[33] Hall P, Fisher NI, Hoffmann B (1994) "On the nonparametric estimation of covariance functions", *Ann. Stat.* 22: 2115-2134.

[34] The International HapMap Consortium (2003) "The International HapMap Project", *Nature*, vol. 426, pp.789–794.

[35] Ho DD , Neumann AU, Perelson AS, Chen W, Leonard JM , Markowitz M(1995) "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection", *Nature*, vol. 373, pp.123–126.

[36] Hou W, Yap JS, Wu S, Liu T, Cheverud JM, Wu RL (2007) "Mapping quantitative trait nucleotides underlying complex traits in a controlled cross", *PLoS ONE* 2(8): e732.

[37] Huang BE, Amos CI, and Lin DY (2007) " Detecting haplotype effects in genomewide association studies", *Genet. Epidemiol.* 31: 803-812.

[38] Huang JZ, Liu N, Pourahmadi M, Liu L(2006) " Covariance selection and estimation via penalized normal likelihood", *Biometrika* 93: 85-98.

[39] Huang Y, Wu H(2006) " A Bayesian approach for estimating antiviral efficacy in HIV dynamic models", *J. Appl. Stat.* 33: 155-174.

[40] Huang Y, Liu D, Wu H(2006) "Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system", *Biometrics* 62: 413-423.

[41] Jin C, Fine JP, Yandell BS (2007) " A unifed semiparametric framework for QTL analyses, with application to spike phenotypes", *J. Am. Stat. Assoc.* 102: 56-67.

[42] Johannes F (2007) "Mapping temporally varying quantitative trait loci in time-to-failure experiments", *Genetics* 175: 855-865.

[43] Kao CH, Zeng ZB, Teasdale RD (1999) "Multiple interval mapping for quantitative trait loci", *Genetics* 152: 1203-1216.

[44] Kitano H (2002) "Computational systems biology", *Nature* 420: 206-210.

[45] Kirkpatrick M, Heckman H (1989) " A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters", *J. Math. Biol.* 27: 429-450.

[46] Kirkpatrick M, Hill WG, Thompson R (1994) "Estimating the covariance structure of traits during growth and aging, illustrated with lactations in dairy cattle", *Genet. Res.* 64: 57-69.

[47] Kloeden PE, Platen E (1999) "Numerical Solution of Stochastic Differential Equations". Springer-Verlag, New York.

[48] Kristensen NR, Madsen H, Jorgensen SB (2004) " Parameter estimation in stochastic grey-box models", *Automatica* 40: 225-237

[49] Lander ES, Bostein D (1989) "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps", *Genetics*, vol. 121, pp.185–199.

[50] Li N, Berg A, Wu RL (2009) " Functional clustering of dynamic gene expression by the ARMA model", *Journal of Biopharmaceutical Statistics* (revised)

[51] Li L, Brown MB, Lee KH, Gupta S (2002) " Estimation and inference for a spline-enhanced population pharmacokinetic model", *Biometrics* 58: 601-611

[52] Li N, Das K, Wu RL (2009) " Functional mapping of human growth trajectories", *J. Theor. Biol.* 260 [Epub of Print].

[53] Li Z, Osborne M, Prvan T (2005) " Parameter estimation in ordinary differential equations", *IMA. J. Num. Anal.* 25: 264-285.

[54] Li HY, Huang ZW, Wu S, Gai JY, Zeng YR, Wu RL (2007) "A conceptual framework to mapping quantitative trait loci governing ontogenetic allometry", *PLoS ONE* 2(8): e1245.

[55] Li Q, Wu RL (2009) "A multilocus model for constructing a linkage disequilibrium map in human populations", *Stat. Appl. Mol. Genet. Biol.*, vol. 8, Article 18.

[56] Liang H, Wu HL (2008) "Parameter estimation for differential equation models using a framework of measurement error in regression model", *J. Am. Stat. Assoc.* 103:1570-1583.

[57] Lin M, Hou W, Li HY, Johnson JA, Wu RL (2007) " Modeling interactive quantitative trait nucleotides for drug response", *Bioinformatics* 23: 1251-1257.

[58] Lin DY, Huang BE (2007) "The use of inferred haplotypes in downstream analyses", *Am. J. Hum. Genet.* 80: 577-579.

[59] Lin DY, Zeng D (2006) "Likelihood-based inference on haplotype effects in association studies (with discussion)", *J. Am. Stat. Assoc.* 101: 89-118.

[60] Little RJA, Rubin DB (2002) *Statistical Analysis with Missing Data*. John Wiley & Son, 2nd, New York.

[61] Liu T, Johnson JA, Casella G, Wu RL (2004) "Sequencing complex diseases with HapMap", *Genetics* 168: 503-511

[62] Luo JT, Hager WW, Wu RL (2009) " A differential equation model for functional mapping of genes affecting short-term virus dynamics", *J. Math. Biol.*

[63] Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, Massachusetts.

[64] Ma CX, Casella G, Wu RL (2002) "Functional mapping of quantitative trait loci underlying the character process: A theoretical framework", *Genetics*, vol. 161, pp.1751–1762.

[65] McLachlan GJ, Krishnan T (2007), "The EM Algorithm and Extensions", John Wiley & Son, 2nd, New Jersey

[66] McLachlan GJ, Peel D (2000), "Finite Mixture Models", John Wiley & Son, New York.

[67] Manichaikul A, Moon JY, Sen S, Yandell BS, Broman KW (2009) " A model selection approach for the identification of quantitative trait loci in experimental crosses", *Genetics* 181: 1077-1086.

[68] Mather K, Jinks JL (1982) " Biometrical Genetics", Chapman and Hall, 3rd Edition, London.

[69] X. L. Meng, D. Rubin (1991) "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm", *Journal of American Statistical Association*, vol. 86, pp.899–909.

[70] X. L. Meng, D. Rubin (1994) "On the global and componentwise rates of convergence of the EM algorithm", *Linear Algebra and its Applications*, vol. 199, pp.413–425.

[71] Animal MK, Meyer K (2000) "Random regressions to model phenotypic variation in monthly weights of Australian beef cows", *Livest Prod. Sci.* 65: 19-38.

[72] Macgregor S, Knott SA, White I, Visscher PM (2005) " Quantitative trait locus analysis of longitudinal quantitative trait data in complex pedigrees", *Genetics* 171: 1365-1376.

[73] Miao H, Dykes C, Demeter LM, Cavenaugh J, Park SY, Perelson AS, Wu H (2008) "Modeling and estimation of kinetic parameters and replicative fitness of HIV-1 from flow-cytometry-based growth competition experiments", *Bull. Math. Biol.* 70: 1749-1771

[74] Miao H, Dykes C, Demeter L, Wu H (2009) "Differential equation modeling of HIV viral fitness experiments: Model identification, model selection, and multi-model inference", *Biometrics* 65: 292-300

[75] Moore JH (2003) "The ubiquitous nature of epistasis in determining susceptibility to common human diseases", *Hum Hered* 56: 73-82.

[76] Moore JH (2005) "A global view of epistasis", *Nat. Genet.* 37: 13-14.

[77] NowakS MA, May RM(2000) *Virus dynamics*. Oxford University Press, New York.

[78] Nettleton D, Doerge RW(2000) "Accounting for variability in the use of permutation testing to detect quantitative trait loci", *Biometrics* 56: 52-58.

[79] Perelson AS (2002) "Modeling viral and immune system dynamics", *Nat. Rev. Immunol.*, vol. 2: 28–36.

[80] Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD (1997) "Decay characteristics of HIV-1-infected compartments during combination therapy", *Nature*, vol. 387, pp.188–191.

[81] Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) "HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time", *Science*, vol. 271, pp.1582–1586.

[82] Pletcher SD, Geyer CJ (1999) "The genetic analysis of age-dependent traits: Modeling the character process", *Genetics* 153: 825-835.

[83] Putter H, Heisterkamp SH, Lange JMA, De Wolf F (2002) "A Bayesian approach to parameter estimation in HIV dynamical models", *Stat. Med.* 21: 2199-2214.

[84] Ramsay JO (1996) "Principal differential analysis: Data reduction by differential operators", *J. Roy. Stat. Soc. Ser. B* 58: 495-508

[85] "Ramsay JO, Silverman BW (2005) Functional Data Analysis (2nd ed.)". Springer, New York.

[86] Ramsay JO, Hooker G, Campbell D, Cao JG (2007) "Parameter estimation for differential equations: a generalized smoothing approach (with discussion)", *J. Roy. Stat. Soc. Ser. B* 69: 741-796.

[87] Redner AR, Walker HF (1984) "Mixture Densities, Maximum Likelihood and the EM Algorithm", *SIAM Review*, Vol. 26, No.2: 195-239.

[88] Ribeiro RM, Bonhoeffer S (2000) "Production of resistant HIV mutants during antiretroviral therapy", *Proc. Natl. Acad. Sci. USA*, vol. 97, 7681–7686.

[89] Robert CP(2001) *The Bayesian Choice*, Springer, 2nd, New York.

[90] Rong L, Gilchrist MA, Feng Z, Perelson AS (2007) "Modeling within-host HIV-1 dynamics and the evolution of drug resistance: trade-offs between viral enzyme function and drug susceptibility", *J. Theor. Biol.*, vol. 247 , 804–818.

[91] Rong L, Gilchrist MA, Feng Z, Perelson AS (2007) "Emergence of HIV-1 drug resistance during antiretroviral treatment", *Bull. Math. Biol.*, vol. 69 , 2027–2060.

[92] Ron M, Weller JI (2007) "From QTL to QTN identification in livestock?winning by points rather than knock-out: a review", *Animal Genetics* 38: 429-439.

[93] Rosenbaum PR, Rubin DB (1983) "The central role of the propensity score in observation studies for causal effects", *Biometrika* 70: 41-55.

[94] Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) "Score tests for association between traits and haplotypes when linkage phase is ambiguous", *Am J. Hum. Genet.* 70: 425-434.

[95] Scheper T, Klinkenberg OD, Pennartz C, van Pelt J (1999) "A mathematical model for the intracellular circadian rhythm generator", *J. Neurosci.* 19: 40-47

[96] Sedaghat AR, Dinoso JB, Shen L, Wilke CO, Siliciano RF (2008) "Decay dynamics of HIV-1 depend on the inhibited stages of the viral life cycle", *Proc. Natl. Acad. Sci.*, vol. 105, pp.4832–4837.

[97] Sen S, Churchill GA (2001) "A statistical framework for quantitative trait mapping", *Genetics* 159: 371-387.

[98] Shankarkumar U, Pawar A, Ghosh K (2009) "HIV-1 evolution, drug resistance, and host genetics: The Indian scenario", *Virus Adaption and Treatment*, vol. 1 , 1–4.

[99] Samoilov MS, Price G, Arkin AP (2006) "From fluctuations to phenotypes: the physiology of noise", *Sci. STKE.* 366: re17.

[100] The STAR Consortium (2008) "SNP and haplotype mapping for genetic analysis in the rat", *Nat. Genet.* 40: 560-566.

[101] Tveito A, Winther R (2009) "Introduction to Partial Differential Equations: A Computational Approach", Springer Verlag, New York.

[102] Varah JM (1982) "A spline least squares method for numerical parameter estimation in differential equations", *SIAM J. Sci. Compt.* 3: 131-141.

[103] Verdugo A, Rand R (2008) "Hopf bifurcation in a DDE model of gene expression", *Commun. Nonlin. Sci. Num. Simul.* 13: 235-242.

[104] Wang NY (2003) "Marginal nonparametric kernel regression accounting for within-subject correlation", *Biometrika* 90: 43-52.

[105] Wang ZH, Hou W, Wu RL (2005) "A statistical model to analyze quantitative trait locus interactions for HIV dynamics from the virus and human genomes", *Stat. Med.*, vol. 25, 495–511.

[106] Wang ZH, Wu RL (2004) "A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics", *Stat. Med.*, vol. 23, pp.3033–3051.

[107] Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, Lifson JD, Bonhoeffer S, Nowak MA, Hahn BH, Saag MS, Shaw Gm (1995) "Viral dynamics in human immunodeficiency virus type 1 infection", *Nature*, vol. 373, pp.117–126.

[108] Whitlock MC, Phillips PC, Moore FB, Tonsor SJ (1995) "Multiple fitness peaks and epistasis", *Ann. Rev. Ecol. Syst.* 26: 601-629

[109] Wodarz D, and Nowak MA (2000) “HIV therapy: Managing resistance”, *Proc. Natl. Acad. Sci. USA*, vol. 97, 8193–8195.

[110] Wolf JB (2000) “Gene interactions from maternal effects”, *Evolution* 54: 1882-1898.

[111] Wu CFJ (1983) “On the Convergence properties of the EM algorithm”, *J. Am. Stat. Assoc.*, Vol. 11: 95-103.

[112] Wu H, Ding A, DeGruttola V (1998) “ Estimation of HIV Dynamic Parameters”, *Statistics in Medicine* 17, 2463-2485.

[113] Wu RL, Lin M (2006) “Functional mapping – How to study the genetic architecture of dynamic complex traits”, *Nat. Rev. Genet.*, vol. 7, pp.229–237.

[114] Wu RL, Lin M (2008) *Statistical and Computational Pharmacogenomics*. Chapman & Hall/CRC, London.

[115] Wu RL, Ma C-X, Casella G (2002) “Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations”, *Genetics*, vol. 160, 779–792.

[116] Wu RL, Ma C-X, Casella G (2007) “ Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL”, Springer-Verlag, New York

[117] Wu RL, Ma C-X, Hou H, Corva P, Medrano JF (2005), “ Functional mapping of quantitative trait loci that interact with the hg gene to regulate growth trajectories in mice”, *Genetics* 171: 239-249.

[118] Wu RL, Ma CX, Lin M, Wang ZH, Casella G (2004) “Functional mapping of growth quantitative trait loci using a transform-both-sides logistic model”, *Biometrics*, vol. 60, pp.729–738.

[119] Wu RL, Ma CX, Lin M, Casella G (2004) “A general framework for analyzing the genetic architecture of developmental characteristics”, *Genetics*, vol. 166, pp.1541–1551.

[120] Wu RL, Ma C-X, Yang MCK, Chang M, Santra U, Wu SS, Huang M, Wang M, Casella G (2003) “Quantitative trait loci for growth in Populus”, *Genet. Res.* 81: 51-64.

[121] Wu WB, Pourahmadi M (2003) “Nonparametric estimation of large covariance matrices of longitudinal data”, *Biometrika* 90: 831-844.

[122] Wu RL, Wang ZH, Zhao W, Cheverud JM (2004) “A mechanistic model for genetic machinery of ontogenetic growth”, *Genetics*, vol. 168, pp.2383-2394.

[123] Wu S, Yang J, Wu RL (2006) "Multilocus linkage disequilibrium mapping of quantitative trait loci that affect HIV dynamics: A simulation approach", *Stat. Med.*, vol. 26, 3826–3849.

[124] Wu S, Yang J, Wang CG, Wu RL (2007) "A general quantitative genetic model for haplotyping a complex trait in humans", *Current Genomics* 8: 343-350.

[125] Xu S (2003) "Estimating polygenic effects using markers of the entire genome", *Genetics* 163: 789-801.

[126] Yang R, Xu S (2007) " Bayesian shrinkage analysis of quantitative trait loci for dynamic traits", *Genetics* 176: 1169-1185.

[127] Yao F, Mller HG, Wang J-L (2005) " Functional data analysis for sparse longitudinal data", *J. Am. Stat. Assoc.* 100: 577-590.

[128] Yao F, Mller HG, Wang J-L (2005) " Functional regression analysis for longitudinal data", *Ann. Stat.* 33: 2873-2903.

[129] Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS (2007) " Efficient strategies for Bayesian mapping of genome-wide interacting QTL", *Genetics* 176: 1865-1877.

[130] Zeidler E (1986) "Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems", Springer-Verlag, New York.

[131] Zeng ZB (1994), "Precision mapping of quantitative trait loci", *Genetics* 136: 1457-1468.

[132] Zens MS, Webb CO (2002) " Sizing up the shape of life", *Science* 295: 1475-1476.

[133] Zhang B, Tong CF, Yin TM, Zhang XY, Zhuge Q, Huang MR, Wang MX, Wu RL (2009) " Detection of quantitative trait loci influencing growth trajectories of adventitious roots in Populus using functional mapping", *Tree Genet. Genom* 5: 539-552.

[134] Zhao W, Chen YQ, Casella G, Cheverud JM, Wu RL (2005) " A nonstationary model for functional mapping of complex traits", *Bioinformatics* 21: 2469-2477.

[135] Zhao W, Hou W, Littell RC, Wu RL (2005) "Structured antedependence models for functional mapping of multivariate longitudinal quantitative traits", *Stat. Appl. Mol. Genet. Biol.*, vol. 4, Article 33.

[136] Zhao W, Ma C-X, Cheverud JM, Wu RL (2004) " A unifying statistical model for QTL mapping of genotype-sex interaction for developmental trajectories", *Physiol. Genom* 19: 218-227.

[137] Zhao W, Zhu J, Gallo-Meagher M, Wu RL (2004) "A unified statistical model for functional mapping of genotype environment interactions for ontogenetic development", *Genetics* 168: 1751-1762.

[138] Zimmerman DL, Nunez-Anton V (2001) "Parametric modeling of growth curve data: An overview (with discussion)", *Test* 10: 1-73.

[139] Zou F, Nie L, Wright FA, Sen PK (2009) "A robust QTL mapping procedure", *J. Stat. Plan Infer.* 139: 978-989".

BIOGRAPHICAL SKETCH

Jiangtao Luo was born in a farm family with a very long history in Chonqing, China. Luos stress the confussian value of study first and working is essential. After years of competitive study, Jiangtao earned his master's degree at Inner Mongolia University in mathematics. His first job was as a lecturer at Chongqing Teachers College, now Chongqing Normal University. Then he pursued further study at Peking University and later taught at Guangzhou Normal University, now named Guangzhou University. Since Jiangtao has been working on his Co-Ph.D. in Mathematics and Statistics but the actual work is concurrent Ph.D. degrees. He has variety of research interests. Right now his main focus is statistical genetics.