

ABSTRACT

GONG, XIAOHUA. Mapping Quantitative Trait Loci in Outbred Half-sib Populations (Under the direction of Professor Zhao-Bang Zeng).

Quantitative trait loci (QTL) mapping in outbred populations faces some challenges unique to the divergent genetic background and complex pedigree relationships. Motivated by a dairy cattle half-sib data set from a grand daughter design, we present in this dissertation a series of endeavors to address various challenges along the analysis flow of QTL mapping. A first step is to infer the haplotypes in sires based on the observed genotypes in sires and his offspring. Our method was shown to outperform peer methods with greater robustness and accuracy yet with fast speed performance. Then in light of adapting the multiple interval mapping method to within-family QTL analysis, we extended the modeling framework by allowing for heteroscedastic residual variances and upgraded the Windows QTL Cartographer accordingly. The advantageous post-analysis result parsing from Windows QTL Cartographer and more importantly, the improved analysis outputs due to more powerful maximum likelihood-based mixture modeling than the least squares regression manifest our efforts in delivering better methodology via practically user friendly software. We further developed a mixed model approach for the purpose of QTL mapping across multiple families that was aimed at modeling QTL effects as both the fixed effect across families and the random effect within families. Our mixed model was shown to encompass similar or higher statistical testing performance on QTL variation than the widely used variance component modeling approach, yet still allowing permutations for obtaining chromosome-wide or genome-wide significance threshold. What's more, the flexibility of our mixed model in constructing alternative hypotheses testing on either fixed or random QTL effects or both was shown to offer interesting insight into the varying sources of signal that would not be unveiled by least squares regression or variance component methods. In concluding our comprehensive approach to QTL linkage mapping in dairy cattle populations, we continue to explore methods of fine mapping by combining both the linkage disequilibrium and linkage information and prospective method improvements are being sought.

Mapping Quantitative Trait Loci in Outbred Half-sib Populations

by
Xiaohua Gong

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics and Statistics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Wenbin Lu

Dr. Steffen Heber

Dr. Zhao-Bang Zeng
Chair of Advisory Committee

Dr. Melissa Ashwell

DEDICATION

To my parents and my wife

BIOGRAPHY

Xiaohua Gong was born in Anhui province in China. He received his Bachelor of Science degree in biology from the University of Science and Technology of China (Hefei, Anhui, China) in 1996. After working at the then-Shanghai Life Science Research Center of the Chinese Academy of Sciences (Shanghai, China) as a research staff for a year, he came to the United States to pursue graduate degrees. He has obtained a Master of Science degree in cell biology from the Albert Einstein College of Medicine of Yeshiva University (Bronx, NY) and a Master of Science degree in information system from the Baruch College of New York (New York, NY). In 2003, he entered the bioinformatics Ph.D. program and pursued a Ph.D. co-major in bioinformatics and statistics at the North Carolina State University (NCSU). Between 2005 and 2008, he also worked at the Discovery Statistics North America of GlaxoSmithKline as a graduate industrial trainee. Prior to and after this internship, he was supported by the genomic science fellowship and teaching assistantship from NCSU.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Zhao-Bang Zeng for his insightful guidance, continuous encouragement and unreserved support to my study and research. I also thank the other committee members: Dr. Melissa Ashwell, Dr. Wenbin Lu and Dr. Steffen Heber, for their insightful advice and precious meetings during the past years. Special thanks go to Dr. Melissa Ashwell for her kind offering the dairy cattle data by which our research was largely motivated, and her support and help during my work and the manuscript writing. I am also grateful to Dr. Shengchu Wang for his great assistance in making upgrades to Windows QTL Cartographer and many of our fruitful discussions.

I want to thank Mr. David Cooper and Dr. Mandy Bergquist for their kind support and guidance during my internship at GlaxoSmithKline (GSK). I also want to thank other members in the Discovery Statistics group and other colleagues at GSK for their numerous suggestions, comments, corrections and assistance on my work. I appreciate the great working experience there.

During my study at the Bioinformatics Research Center (BRC) and the Department of Statistics (StatDept), I received a lot of help from many people and I thank them all for making all the resources available for me, without which I would not be able to complete my study. I am indebted to Juliebeth Briseno, Tina Chen and Adrian Blue for their help in dealing with various paperwork. Chris Smith and Stanton Martin gave me much assistance on IT-related issues and I appreciate their help very much. I would also thank other faculty members at the BRC and the StatDept, Dr. Helen Zhang, Dr. Daowen Zhang and Dr. Eric Stone, et al., for their kind and helpful advice. I thank my fellow students and friends at BRC and StatDept for their various kinds of help and knowledge sharing, etc.

Lastly, I thank my parents and my elder brother for their deep love, continuous support and heartwarming encouragement throughout these years. In particular, I thank my lovely wife, Youfang Liu, for her sincere and endless love. Without her tolerance and generous support, I would not be able to come to where I am today.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS	x
1 Introduction	1
1.1 A General View of QTL Mapping	1
1.2 Population, Design and Marker Analysis	2
1.2.1 Inbred backcross designs and its marker analysis	4
1.2.2 Outbred designs, linkage phase and IBD	5
1.3 QTL Analysis Methods (in Outbred Populations)	12
1.3.1 Linear fixed effects models	13
1.3.2 Linear random effects model	18
1.3.3 Bayesian QTL mapping	22
1.3.4 Nonparametric QTL mapping	27
1.4 Background and Motivations of Our Research	28
1.4.1 Genetic evaluation of dairy traits	28
1.4.2 Dairy cattle data for QTL mapping analysis	31
1.4.3 Motivating data sets	32
1.4.4 Outline of dissertation	32
2 Haplotyping in Outbred Half-sib Populations	35
2.1 Introduction	35
2.2 Review of Selected Haplotyping Methods	37
2.2.1 Knott et al. (1996) method	37
2.2.2 Georges et al. (1995) method	38
2.2.3 Large scale haplotype phaser - LSPH (2006)	40
2.2.4 Windig-Meuwissen (2004) method	41
2.3 Our Method for Inferring Sire Haplotypes	43
2.3.1 Reconstruction of sire haplotypes	44
2.3.2 Inference procedures of most probable sire haplotypes	46
2.4 Simulations	48
2.5 Results	49
2.6 Discussions	53
2.7 Future Work	56
2.7.1 Speed performance improvement	56

2.7.2	Extension to mix sibship	57
2.7.3	Inference of offspring haplotypes	58
3	Multiple Interval Mapping in Large Half-sib Families	59
3.1	Introduction	59
3.2	Single Half-sib Family QTL Analysis: Least Squares Method	61
3.3	Our Method: Weighted Multiple Interval Mapping (wMIM)	63
3.3.1	Maximum likelihood analysis	64
3.3.2	Hypothesis testing	66
3.3.3	Calculation of R^2	67
3.3.4	Windows QTL Cartographer software updates	68
3.4	Application of wMIM to Experimental Data	68
3.4.1	Data explorative analysis	69
3.4.2	Within-family QTL analysis results	77
3.4.3	Discussions	80
3.5	Prospectives	82
3.5.1	Further upgrades to Windows QTL Cartographer	82
3.5.2	Multiple family QTL analysis as fixed effects	83
4	Across-family Mixed Model Analysis of QTL	84
4.1	Introduction	84
4.2	Variance Components Approach	86
4.3	Our Approach: Mixed Effects Model	88
4.3.1	Mixed effects model specifications	88
4.3.2	Hypothesis testing	89
4.3.3	Calculation of R^2	91
4.4	Simulation Studies of Mixed Model Approach	91
4.4.1	Type I error for across-family analysis: pedigree versus unrelated	92
4.4.2	Choice of LOD-drop intervals	93
4.5	Application to Experimental Data	97
4.5.1	Our mixed model analysis	97
4.5.2	Variance components analysis	97
4.5.3	Mixed model analysis results	99
4.5.4	Comparison with variance components analysis results	103
4.6	Discussions	103
5	Summary and Discussions	107
5.1	Summary and Discussions	107
5.2	Towards Fine Mapping	109
5.2.1	Linkage disequilibrium linkage analysis - LDLA	110
5.2.2	Variant methods and prospective extensions	112

APPENDICES	125
A Multiple Family Multiple Interval Mapping.....	126
A.1 Model and Notations	126
A.2 Complete Data Likelihood	127
A.3 E-M Algorithm	128
A.3.1 The $[s + 1]$ -th E-step:	128
A.3.2 The $[s]$ -th M-step:	129
A.3.3 REL is available	130
A.4 Discussions About QTL Testing	130
B Approximation of Mixed-Mixture Model into Mixed Model.....	132
B.1 Mixed Effects Mixture Model	132
B.1.1 Model specifications and notations	133
B.1.2 Mixture of variance structures	135
B.1.3 Complete data likelihood	136
B.2 Approximated Mixed Effects Model	137

LIST OF TABLES

Table 1.1	Conditional probabilities of QTL genotypes in backcross design.	6
Table 1.2	Kinship coefficients for simple relative relationships.	19
Table 2.1	Speed performances of sire haplotype reconstruction methods.	51
Table 3.1	Marker map summary of the chromosomal segment on BTA 18.	70
Table 3.2	Microsatellite marker data summary.	71
Table 3.3	Phenotypic data summary.	75
Table 3.4	Summary of suggestive QTL from within-family interval mapping. ...	78
Table 4.1	Type I error rates for treating pedigree-based families as unrelated. .	94
Table 4.2	Confidence levels of various LOD-drop intervals.	97
Table 4.3	Across-family mixed model analysis results.	100
Table 5.1	Haplotype-based IBD coefficient matrix at a locus p	111

LIST OF FIGURES

Figure 1.1	Diagram of grand daughter design in dairy cattle.	29
Figure 2.1	Deduction of sire gametes in the offspring.	47
Figure 2.2	Speed performance comparison of haplotype reconstruction methods.	50
Figure 2.3	Accuracy of haplotype reconstruction.	52
Figure 2.4	Venn diagram of best-accuracy-achieving methods.	54
Figure 3.1	Diagram for calculating conditional probability.	64
Figure 3.2	Information content for the marker coverage region for each family.	73
Figure 3.3	Box plot of each trait within each family.	74
Figure 3.4	Reliability of different traits.	76
Figure 3.5	QTL interval mapping results.	79
Figure 4.1	Across-family analysis results.	101
Figure 4.2	Decomposition of likelihood ratios in our mixed model analysis.	102
Figure 4.3	Comparison of likelihood ratio profiles on QTL random effects only.	104
Figure B.1	Diagram of mixed QTL effects in multiple family populations.	133

LIST OF ABBREVIATIONS

AIPL	Animal Improvement Program Laboratory
CE	Calving Ease
cM	Centi-Morgan
DD/DYD	Daughter (Yield) Deviation (of bulls)
DPR	Daughter Pregnancy Rate
EM	Expectation-Maximization (Algorithm)
FP	Fat Percentage
FY	Fat Yield
GDD	Grand Daughter Design
IBD	Identity/Identical By Descent
IBS	Identity/Identical By State
IM	Interval Mapping
LD	Linkage Disequilibrium
LOD	Logarithm of Odds
LR/LRT	Likelihood Ratio (Test)
LS	Least Squares
LSPH	Large Scale Phase Haplotyper
MARC	Meat Animal Research Center
MCMC	Monte Carlo Markov Chain
MIM	Multiple Interval Mapping
ML/MLE	Maximum Likelihood (Estimate)
MY	Milk Yield
PDB	Percent of Difficult Births
PL	Productive Life
PP	Protein Percentage
PTA	Predicted Transmitting Ability
PY	Protein Yield
QTL	Quantitative Trait Locus (or Loci)
REL	Reliability
REML	Restricted Maximum likelihood
SCS	Somatic Cell Score
SNP	Single Nucleotide Polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
USDA	United States Department of Agriculture
VC/VCM	Variance Components (Model)
WinQTLCart	Windows QTL Cartographer
wMIM	Weighted Multiple Interval Mapping

Chapter 1

Introduction

In this chapter, we review background literature that is relevant with analyzing quantitative trait loci (QTL) in outbred half-sib populations, the main theme of this dissertation. We first present the QTL mapping problem in a general formulation. Then we introduce commonly used population designs, inbred and/or outbred, and their respective features. Next we review both the genetic marker analysis and QTL modeling parts for QTL mapping analysis, with emphasis on the statistical aspects related with outbred populations. In the last section of this chapter, we outline our own research and the organization of the remaining chapters of this dissertation.

1.1 A General View of QTL Mapping

Quantitative genetics is concerned with the inheritance of quantitative traits between individuals. Genetic studies in quantitative traits can be on morphological traits, physiological and biochemical traits, complex genetic diseases and behavioral traits, etc. Genes that affect quantitative trait variation in a population are called quantitative trait loci. QTL analysis is to localize QTL on a genetic linkage map such as number and genomic positions of QTL, as well as to characterize genetic architecture of QTL such as the effects of QTL alleles within (dominance) and between (epistasis) loci, pleiotropic effects of QTL and QTL-by-environment interactions, etc.

From a likelihood-based point of view, QTL mapping analysis is to link observed

trait phenotypes (\mathbf{Y}) to unobserved QTL genotypes ($\boldsymbol{\theta}$) as well as other non-genetic factors, if any, through observed genetic marker phenotypes (\mathbf{X}). That is, the likelihood of data for a genetic model can be written in the following symbolic form [ZENG, 2000]:

$$L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) = \Pr(\mathbf{g}|\mathbf{X}, \lambda)f(\boldsymbol{\theta} \mapsto \mathbf{Y}|\mathbf{g}) \quad (1.1.1)$$

where

- \mathbf{g} specifies the genotypes of QTL.
- $\boldsymbol{\theta} \mapsto \mathbf{Y}$ specifies the mapping from QTL genotypes to trait phenotypes (*a.k.a.*, the model) and $f(\boldsymbol{\theta} \mapsto \mathbf{Y}|\mathbf{g})$ is a function of the mapping given the QTL genotypes.
- $\Pr(\mathbf{g}|\mathbf{X}, \lambda)$ specifies the probability of QTL genotypes at specified genomic locations (λ) given the observed genetic marker phenotypes.

Thus, QTL mapping analysis consists of two parts as shown above: (i) the marker analysis part that involves QTL genotypes inference given the observed marker phenotypes at various genomic locations for various experimental designs and data structures; (ii) the model selection and evaluation part of the genotype-phenotype mapping relationship for the quantitative trait(s). That is, we can separate the analysis of these two parts and study them independently.

1.2 Population, Design and Marker Analysis

Linkage refers to the status of deviation from free recombination between two loci and a linkage map, also called a genetic map, refers to the order and spacing of markers in which the genetic distance is used. The distance measure used in a linkage map uses Morgan as the unit that is defined as the expected number of recombination between two loci over one meiosis. Centimorgan (cM), one hundredth of a Morgan, is more frequently used. A genetic linkage map has no explicit relationship with a physical

map in which the order and spacing of markers are based on the actual number of nucleotides, the molecular building blocks of genes, chromosomes and genomes.

As outlined in the general formulation in equation (1.1.1), the marker analysis of QTL mapping involves inferring the unobserved QTL genotypes based on the observed genetic marker phenotypes by appropriately dealing with various experimental designs and data structure. A complete approach of marker analysis may include the linkage map analysis such as grouping markers into different linkage groups and ordering markers within the same linkage group. However, in this dissertation, we will assume that a linkage map of genetic markers is known because our focus of marker analysis is on the inference of linkage phase and QTL genotypes. Furthermore, we consider populations from biparental crosses for diploid organisms only.

There exist many population structures that are used for QTL mapping analysis in medical research, animal improvement and plant breeding, etc. We can divide them into two main categories: inbred and outbred populations. The outbred populations can be further divided such as outbreeding within or between domesticated species, outbreeding within or between natural undomesticated unmanipulated populations, and outbreeding between a domesticated population and an undomesticated population, etc. We will confine our discussion to outbreeding within livestock.

In the genetics literature, an inbred line (or strain) is a set of genetically related individuals whose chromosomes are homozygous pairs and at any given gene the maternal and paternal alleles are identical. Such inbred lines are usually obtained by repetitively mating within the same pool of line (or selfing) over multiple (8 - 10 or more) generations, during which process the expected number of heterozygous loci is halved each generation. Inbred lines are typical of plants such as maize and rice, and some model animals such as mice, rats, chickens, etc. For many other species, especially for most of animals, such as apple, pine, cattle, horse, etc., inbred lines are not practically feasible due to various reasons ranging from too much resources needed (time and money), severe inbreeding depression and immorality (for instance, human populations).

For the purpose of QTL mapping analysis, such species are sampled/collected in the form of pedigree, with individuals over multiple generations being collected and

genotyped/phenotyped. A pedigree is a pictorial description of genealogical structure among a set of individuals [THOMPSON, 2000]. It consists of nuclear families (parent and offspring). Individuals that have no parents in a pedigree is called founder and all other individuals are non-founders.

In QTL mapping analysis, the gene(s) is unknown and unobserved thus genetic markers that have different variants in parents are sought and the marker phenotypes are measured in parents and offspring for the purpose of QTL genotype inference. Microsatellite (or simple sequence repeat, SSR) and single nucleotide polymorphism (SNP) markers are the two main types of genetic markers nowadays. Both microsatellite and SNP are co-dominant markers. A microsatellite marker is more informative than SNP marker in that it is usually poly-allelic, however, multiple SNP markers in combination can overcome this shortfall. On the other hand, the bi-allelic SNP markers can be typed at a much higher density (as small as dozens or hundreds of nucleotides apart) in a very high throughput manner that makes it possible to finely locating QTL to specific candidate genes. In this dissertation, we mainly refer to microsatellite marker data for QTL mapping analysis, though the theory and methods are applicable towards SNP marker data, too.

1.2.1 Inbred backcross designs and its marker analysis

All inbred designs starts with two inbred parental lines, say P_1 and P_2 , that differ in the trait of interest and such contrast of trait phenotypes is due to the different variants of some genetically heritable components (genes) in parents. Within each line all individuals are genetically identical and homozygous at all loci. When the individuals from the two parental strains are intercrossed, the F_1 offspring will be obtained that are genetically identical with one another but heterozygous for all loci where the two parental lines differ. Depending on the various mating strategies of using F_1 offspring in the next step, several kinds of inbred populations can be obtained such as F_2 , backcross and recombinant inbred lines, etc. However, with respect to the main focus of this dissertation on the outbred half-sib design, we will only describe in detail the inbred backcross design below.

Backcross populations is obtained by crossing F_1 offspring with one of the two parental lines, say P_1 . Because P_1 is homozygous at every locus (say QQ) and that F_1 is heterozygous at every locus (Qq) with one allele (Q) from P_1 and another allele (q) from P_2 , the backcross offspring would bear at any locus Q , by a matter of chance, one of the two genotypes: QQ or Qq . Obviously, the backcross of F_1 with P_2 would give another backcross population in which the two possible genotypes are Qq and qq . For simplicity, we always refer to backcross as the cross between F_1 and P_1 throughout this dissertation unless otherwise specified.

The marker analysis of an inbred backcross population is relatively simple because the marker linkage phases in the mating partners, the P_1 and F_1 individuals, are known of certainty. The only remaining task is to infer the unobserved QTL genotypes in a probabilistic manner given the observed markers that have linkage with the QTL locus. Such formulas are readily available (Table 1.2.1) for simple cases where a single QTL sits in an interval of two flanking markers (denoted as M and N) with the ordering known as MQN and the recombination frequencies also known as r_{MQ} between M and Q , r_{QN} between Q and N , and r_{MN} between M and N , respectively. In cases where genotypes of the closest flanking markers may be missing due to various causes such as genotyping failures, the probabilities of QTL genotypes may be inferred by using the closest fully informative neighboring markers through a Markov chain method [JIANG and ZENG, 1997].

1.2.2 Outbred designs, linkage phase and IBD

Outbreeding refers to cross between two unrelated or distantly related parent lines. The two parent lines are not RILs, so there is no guarantee that parental lines are homozygous at all loci. For co-dominant markers such as microsatellites, there could be a maximum of four different alleles observed in the two parents within a cross, if not considering the null allele. The number of marker allele variants could vary between 1, 2, 3 or 4 from one locus to another on a chromosome, among which the marker showing four allele variants are the most informative. The genotypes in

Table 1.1: Probabilities of QTL genotypes conditional on flanking marker genotypes in backcross design. Note that $\gamma = r_{MQ}/r_{MN}$, that is, the relative position of QTL with respect to the two flanking markers M and N .

Marker Genotypes		QTL Genotypes	
		QQ	Qq
MM	NN	$\frac{(1 - r_{MQ})(1 - r_{QN})}{1 - r_{MN}} \approx 1$	$\frac{r_{MQ}r_{QN}}{1 - r_{MN}} \approx 0$
MM	Nn	$\frac{(1 - r_{MQ})r_{QN}}{r_{MN}} \approx 1 - \gamma$	$\frac{r_{MQ}(1 - r_{QN})}{r_{MN}} \approx \gamma$
Mm	NN	$\frac{r_{MQ}(1 - r_{QN})}{r_{MN}} \approx \gamma$	$\frac{(1 - r_{MQ})r_{QN}}{r_{MN}} \approx 1 - \gamma$
Mm	Nn	$\frac{r_{MQ}r_{QN}}{1 - r_{MN}} \approx 0$	$\frac{(1 - r_{MQ})(1 - r_{QN})}{1 - r_{MN}} \approx 1$

offspring from an outbreeding cross can be highly divergent that makes it easy to trace the paternal or maternal alleles, which in itself could be an advantageous feature of outbreeding strategy.

Outbred designs

Commonly seen outbred designs include half-sib, full-sib and mix-sib, etc. There is an analogy between outbred half-sibs and full-sibs with the inbred backcross and F_2 populations, respectively, though fundamental differences exist. The inbred F_1 individuals are heterozygous at all markers and QTL that are of different variants between the two parental RILs. The linkage phase between markers and QTL is identical in all inbred F_1 's. For outbred populations, however, since there is no guarantee that every parent is heterozygous at both marker loci and QTL, QTL segregation in any a family may be uncertain. For instance, assuming a bi-allelic QTL with equal allelic frequencies, half of the parents are expected to be homozygous at this QTL under random mating thus non-informative with respect to the purpose of detecting it. A second complication kicks in as the linkage phases between markers and QTL vary across outbred families, due to independent crossing-over events during gametogenesis. Therefore, QTL effects are either considered within each family as fixed effects (*a.k.a.*, a difference in means of different genotypes) or as random effects (proportion

of the total phenotypic variance explained by QTL, *a.k.a.*, QTL heritability). A third but not the last issue with outbred populations is that there may not be sufficiently large sibships for QTL mapping. In the contrast, all sibships from an inbred cross can be regarded as a single large family.

Choice of an outbred design for studying a species depends on many factors such as practical feasibility and manipulability, resources needed (time, money, labor cost, etc.), etc. Full-sib designs, in which offspring share both of the parents, are commonly seen for pine, for example, because there can be hundreds of seeds (full-sibs) per cross. In the contrast, half-sib-based designs are used for cattle populations, in which offspring share one of the two parents (usually the bull, thus called paternal half-sibship). Reasons for choosing half-sib designs for cattle are many with one being that a cow usually gives birth to only one offspring per litter and the pregnancy takes one year or so. Further complicate sibships arise where multiple full-sib crosses may share one or the other parents among them. For example, the litter sizes are usually around 10 for pigs and a few dozen for chickens. These sizes may not be sufficiently large by themselves for QTL detection, but by sharing the common male(s), a much larger population consisting of hundreds of mix-sibs can be obtained in which offspring are of full-sibship within the same cross but are half-sibs between two crosses.

Linkage phase

Given that a marker linkage map is assumed to be known, marker analysis in outbred designs need to first tackle the challenge of inferring the linkage phases, which, as discussed previously, is not a problem at all for inbred designs. The inference of linkage phase, also called haplotyping, refers to the reconstruction of haplotypes from observed unordered marker genotypes within a pedigree. Current genotyping technology cannot offer experimentally-determined ordered marker genotypes (*a.k.a.*, molecular haplotyping) at a high throughput pace and in a cost-effective way [JUDSON and STEPHENS, 2001]. It would help reduce the ambiguity in phasing if close relatives are also genotyped, however, the haplotyping could still become problematic when the number of loci is only moderately large. Missing values in genotype data also increase

the challenge for haplotyping. Thus, *in silico* haplotyping becomes an important computation task.

Various algorithms and methods exist for different scenarios such as the pedigree complexity, the marker density and the chromosomal segment length of marker coverage, etc. These methods can be grouped into population-based or pedigree-based according to the sampling strategy and the study designs. Within the context of QTL mapping analysis in outbred populations, we are concerned with pedigree-based methodologies here. Then again the methods can be divided into two categories: some methods are heuristic rule-based [WIJSMAN, 1987; O’CONNELL, 2000; TAPADAR *et al.*, 2000; QIAN and BECKMANN, 2002; LI and JIANG, 2003, 2004; ZHANG *et al.*, 2005] and some are statistical likelihood-based [ELSTON and STEWART, 1971; LANDER and GREEN, 1987; WEEKS *et al.*, 1995; SOBEL and LANGE, 1996; KRUGLYAK *et al.*, 1996; LIN and SPEED, 1997; DU *et al.*, 1998; THOMAS *et al.*, 2000; ABECASIS *et al.*, 2002]. However, there is no guarantee for a method to find the most probable haplotype configurations, especially when handling a large number of loci because the too large number of consistent haplotype configurations usually forbids exhaustive searches [LIN and SPEED, 1997].

The rule-based methods iteratively deduce marker linkage phases in parents and offspring in accordance with the Mendelian rules of genetics and common logical rules. This requires sufficient information about the genotypes in parents and offspring, and if possible, their relatives. The detail logical flow of applying the Mendelian rules can be lengthy, iterative and complex. If the marker density is high (such as microsatellite or SNP) and the chromosomal segment is not known to contain recombination hot-spots, certain biological assumptions such as none or a minimal number of recombination crossing-overs during gametogenesis are usually undertaken. Rule-based methods are *ad hoc*, deterministic and usually fast. However, they do not make full use of marker information such as the inter-marker distances as well as missing or uninformative marker data, in which parent(s) and offspring have identical heterozygous genotypes. The widely used assumption of no recombination could also be violated in regions containing recombination hot-spot(s) even for highly dense SNP markers.

In the likelihood-based category, appropriate likelihood models of linkage phases given the observed un-phased genotypes are constructed and optimization of such likelihood function gives the most probable linkage phase(s). Approximated likelihood and/or conditional likelihoods are also used to reduce the computational or algorithmic complexities. Such likelihoods are typically stochastic [SOBEL and LANGE, 1996; LIN and SPEED, 1997; THOMAS *et al.*, 2000] and demand great computation resources (computing time and memory and storage capacities) in the face of large pedigrees, and/or large number of markers and/or in the presence of missing/uninformative data, etc. There are also efforts of developing deterministic (approximated) likelihood-based methods for haplotyping, such as DU *et al.* [1998]; GAO *et al.* [2004]; GAO and HOESCHELE [2008], that may offer faster computation speed.

Identity by descent (IBD)

Once the linkage phases are inferred, or a set of linkage phases are inferred with respective probabilities, the pairwise identity-by-descent (IBD) coefficients between two individuals at a putative QTL position are needed for QTL mapping analysis. IBD refers to such a state of two alleles at a single locus, be they within the same individual or between individuals, that are identical copies of the same allele in some earlier generations. In other words, both alleles are copies that arise by DNA replication from the same ancestral sequence without any intervening mutations.

Identical allele type between two alleles, as called IBS (identity-by-state), is a necessary but not the sufficient condition for IBD, because the same allele type could have transmitted from different ancestors. Between any two diploidy individuals, the IBD status for a locus can only be one of the three values: 0, 1 or 2 alleles IBD, should we know the exact ancestry of each allele in the two individuals. We denote a vector, (k_0, k_1, k_2) , to represent the probabilities of IBD status equal to 0, 1 or 2, respectively, and then based on relationship only and under the assumption of no inbreeding, we can set the prior probabilities in the vector above. For instance, the vector of prior probabilities is $(1, 0, 0)$ for unrelated pair of individuals, $(0, 1, 0)$ for parent-offspring pair, $(0, 0, 1)$ for monozygotic twins and $(1/4, 1/2, 1/4)$ for a sib

pair. Such prior IBD probabilities leads to another important concept, the kinship coefficient (the coefficient of co-ancestry) that measures the chance of a randomly chosen allele from one individual being IBD to a randomly chosen allele from a second individual. Easy to see, the kinship coefficient for a pair of non-inbreeding individuals can be obtained by the equation $kinship = k_1/4 + k_2/2$. For the simple relationships aforementioned, their kinship coefficients are 0, 1/4, 1/2 and 1/4 for unrelated, parent-offspring, monozygotic twins and sib pairs, respectively.

With marker genotype data, the posterior IBD probabilities can be computed at a marker locus or any genomic loci based on the observed marker IBS status in pedigrees. This has been a challenging task and many methods and their improvements have been proposed in the past two decades. Roughly, these methods differ in terms of algorithm complexity, computational cost, robustness in applicability and etc., due to their differing problem specifications such as the size and complexity of pedigrees, the number and density of genotyped markers, (partially) missing marker genotypes or not in founders or parents, etc.

A general recursive formula for calculating IBD probabilities between two alleles (Q_i^k and $Q_{i'}^{k'}$), in which $k = 1$ or 2 denotes the paternally (1) or maternally (2) inherited allele and i and i' are two individuals who are not of parent-offspring relationship, is given below:

$$\begin{aligned}
 \Pr(Q_i^1 \equiv Q_{i'}^{k'}) &= \Pr(Q_i^1 \leftarrow Q_f^1 | \mathbf{M}) \Pr(Q_f^1 \equiv Q_{i'}^{k'} | \mathbf{M}) \\
 &+ \Pr(Q_i^1 \leftarrow Q_f^2 | \mathbf{M}) \Pr(Q_f^2 \equiv Q_{i'}^{k'} | \mathbf{M}) \\
 &+ \Pr(Q_i^1 \leftarrow Q_m^1 | \mathbf{M}) \Pr(Q_m^1 \equiv Q_{i'}^{k'} | \mathbf{M}) \\
 &+ \Pr(Q_i^1 \leftarrow Q_m^2 | \mathbf{M}) \Pr(Q_m^2 \equiv Q_{i'}^{k'} | \mathbf{M}). \quad (1.2.1)
 \end{aligned}$$

where \mathbf{M} denotes observed marker information, \equiv denotes IBD relationship, \leftarrow denotes the transmission of an allele from parent to offspring and the subscripts f or m are father or mother. There are obvious constraints such as IBD coefficient is 1 between one and oneself or monozygotic twins, and 0 between unrelated individuals.

Each term in this equation can be further recursed according to the pedigree and

Mendelian rules of genetics. Such recursions were discussed in WANG *et al.* [1995] and DAVIS *et al.* [1996] in more general scenarios such as IBD of QTL with a single or multiple linked genotyped markers. Note that if the marker linkage phase of all or some individuals can be inferred with certainty, assuming no genotyping errors, the recursions steps will be simplified. Exact enumerations for the recursive peeling procedure will become clumsy when dealing with large and complex pedigrees, for which Monte Carlo algorithms are of great uses [DAVIS *et al.*, 1996; HEATH, 1997].

Software for IBD estimation

There are many methods proposed for IBD computation and we will cherry-pick a few that have been implemented into software programs and are easily accessible:

Loki implements a Markov Chain Monte Carlo (MCMC) algorithm for computing multipoint IBD probabilities [HEATH, 1997]. It is suitable for large and complex pedigrees, however, the computation cost would also be high: long running time due to the computational complexity and the need for computing resources for multiple independent Markov chains to diagnose the MCMC convergence. Loki software is available at <http://www.stat.washington.edu/thompson/Genepi/Loki.shtml> and it runs on Unix or Linux operating system.

Merlin implements a divide-and-conquer algorithm [ABECASIS *et al.*, 2002], first proposed by INDURY and ELSTON [1997], with incorporation of a sparse binary inheritance tree representing gene flow pattern that is much improvement over the full inheritance tree adopted by Lander and Green algorithm [LANDER and GREEN, 1987]. Merlin is available at <http://www.sph.umich.edu/csg/abecasis/Merlin/> and can be used in both Windows and Unix/Linux operating systems.

SOLAR (sequential oligogenic linkage analysis routines) implements a two-step procedure for IBD computation [ALMASY and BLANGERO, 1998]. In the first step, SOLAR uses either a Monte Carlo algorithm (the default method) due to DAVIS *et al.* [1996] or the Curtis and Sham algorithm [CURTIS and SHAM, 1995]

for marker-specific IBD. In the second step, SOLAR implements a regression-based method [ALMASY and BLANGERO, 1998], which was an extension of FULKER *et al.* [1995] approximate multipoint calculations for sib pairs, for calculating multipoint IBD. SOLAR also supports the use of other software for IBD computation such as SimWalk2 [SOBEL and LANGE, 1996], Loki, GeneHunter [KRUGLYAK *et al.*, 1996] and Merlin. SOLAR software is available at <http://solar.sfbgenetics.org/> and requires Unix or Linux operating system.

1.3 QTL Analysis Methods (in Outbred Populations)

In linkage analysis of outbred populations that was formed recently, distance of QTL and markers are usually of moderate resolution (*e.g.*, a 10cM map) such that it is expected to be in linkage equilibrium between markers and QTL [HOESCHELE, 2007]. In such cases, linkage disequilibrium caused by ancestral mutations will have decayed to zero over time. That is to say, in QTL linkage analysis of outbred pedigrees, QTL effects have to be estimated within parents [HOESCHELE, 2007].

Methods of QTL analysis have been continuously improved for higher power and more complex designs. As the quest to understanding complex genetic architectures increases, the methods themselves become more and more statistically and computationally involved. In general, outbred designs have more complex population structures and the more complex the data structures are, the more complex the analytic methods are needed. A single large half-sib or full-sib family could find many of the QTL analysis methods used for inbred lines applicable with appropriate adaptations; whereas study designs with multiple pedigrees over multiple generations would call for much more sophisticated statistical modeling and analysis.

A variety of statistical QTL mapping methods have been developed and we can roughly catalog them into linear models and generalized linear models, Bayesian approach, non-parametric and semi-parametric. Linear models are a major statistical

QTL modeling framework ranging from linear regression on single markers to multiple markers, from single QTL interval mapping to multiple QTL multiple interval mapping, from single trait analysis to joint analysis of multiple traits, with or without epistasis, etc. In their formulations, the (transformation of) trait phenotypes are the response variables (\mathbf{Y}) and marker/QTL genotype data alongside with other covariates are explanatory variables (\mathbf{X}), with the residuals usually assumed to be independent identical normally distributed.

1.3.1 Linear fixed effects models

Methods described in this part are all variants of linear models in which effects of all covariates and QTL/marker are fixed, leaving the residual to be the only variance component. Such fixed effects modeling is simple in terms of both theory and numerical implementations, especially when compared with the random effects model.

Linear regression on markers

Linear regression on markers for mapping QTL was started as early as 1923 and continuous attention has been paid to this subject in many years that followed (see review in DOERGE *et al.* [1997]). The linear model is usually specified as follows with respect to an inbred backcross:

$$y_j = \mu + bx_j + e_j, \quad j = 1, 2, \dots, n \quad (1.3.1)$$

where

- y_j is the observed trait phenotype of j -th individual;
- μ is the overall mean of the trait value;
- b is the additive marker effect;
- x_j is the marker genotype;

- e_j is the residual following a normal distribution with mean zero and variance σ_e^2 .

Its basic assumption is that if some markers are close to the QTL (*a.k.a.*, linked), these two loci tend to be transmitted together. One unique feature for this method is that it does not require a genetic map *per se*. The essence of regression analysis on markers is to test for significant differences between or among the trait phenotypic means of different marker genotypes. Such a method was for inbred lines where the allele segregation in progeny is easy to determine most of time, but is also adaptable to certain types of outbred families, in which if marker allele transmission from parents to offspring can be successfully traced and the sample sizes for different groups of marker genotypes are large enough. Significant differences in mean phenotypes between or among different groups of parental marker alleles would reveal the association between marker(s) being tested and a linked QTL nearby. However, there would be no clue where the putative QTL lie and its effects. This type of analysis has only very limited usage for linkage analysis now due to the reasons aforementioned and in addition, its inability of handling complex population structures and missing or non-informative markers, and low power performance in detecting QTL signal.

Single QTL models

One of the most influential method for mapping a single QTL can be credited to LANDER and BOTSTEIN [1989]. The Lander-Botstein interval mapping (IM) method takes up a linear regression model with one QTL postulated to occur somewhere in the genome. The linear model is usually specified as follows with respect to an inbred backcross:

$$y_j = \mu + b^*x_j^* + e_j, \quad j = 1, 2, \dots, n \quad (1.3.2)$$

where

- y_j is the observed trait phenotype of j -th individual;
- μ is the overall mean of the trait value;

- b^* is the additive QTL effect;
- x_j^* is the unobserved QTL genotype;
- e_j is the residual following a normal distribution with mean zero and variance σ_e^2 .

Note the difference in model specifications between interval mapping (equation 1.3.2) and marker regression (equation 1.3.1): x_j is observed marker genotypes while x_j^* is unobserved QTL genotype.

The unobserved QTL genotypes can be inferred from the observed marker data, as demonstrated with inbred backcross designs in Table 1.2.1, should the marker linkage map be available and QTL location be known. To overcome the difficulty that the QTL location is unknown, Lander-Botstein's linear model steps through marker intervals of small increments (say 1 cM) and the likelihood ratio test (LRT, or its equivalent LOD test) for QTL is conducted and the maximum log likelihood ratios over the scanned positions would give the strongest evidence for QTL existence. By this way, the IM method would detect a single QTL if the test statistics at some scanned position exceed some specified threshold value. LANDER and BOTSTEIN [1989] showed that the distribution of the LRT statistic approximates an Ornstein-Uhlenbeck diffusion process under the null model with no QTL on the chromosome. In practice, however, the permutation procedures proposed by CHURCHILL and DOERGE [1994] and DOERGE and CHURCHILL [1996] are widely used for obtaining chromosome-wide (or genome-wide) significance threshold.

By scanning through a series of putative positions, Lander-Botstein IM also avoids simultaneous estimation of the location and effects parameters of QTL thus simplifying the likelihood maximization. The maximum likelihood procedures aforementioned are usually implemented by an expectation-maximization (EM) algorithm because the Lander-Botstein linear model is actually a mixture models at every postulated QTL location. The detailed derivations of the EM algorithms will be omitted here, however, because we will present much of it in detail in an extension of the mixture model-based multiple interval mapping in outbred half-sib cattle populations in

Chapter 3.

The IM was a breakthrough from single marker analysis in that both locations and effects of QTL can be estimated; however, its assumption of one or zero QTL on a chromosome can be easily violated and nearby QTL could affect significantly the current interval of interest, which causes biases in estimated position and effects of the current QTL.

A least squares-based approximation of the IM was proposed by MARTINEZ and CURNOW [1992], HALEY and KNOTT [1992] and HALEY *et al.* [1994] that takes the similar form of linear regression model except that the unknown QTL genotypes being replaced by its probability (alternative parameterizations exist but they are statistically equivalent with respect to testing for the existence of QTL). Testing and estimation from such models become an ordinary linear regression problem and standard least squares analysis can be easily conducted using general purpose statistical software such as SAS (www.sas.com) or R (www.r-project.org). To address the problem of the unknown location of QTL, such least squares model can also step through a series of putative positions, as IM does, and the profile of test statistics (F-statistic) over the scanned genomic region will be examined for evidence of whether and where a putative QTL may be located. In literature, this least squares-based interval mapping is also called regression interval mapping (RIM).

The advantage of the least squares method is its outstanding computational efficiency. It also allows easy model specification changes by adding or dropping covariates without demanding customized modifications to the model evaluation, testing and estimation procedures. In the contrast, the maximum likelihood analysis of IM could be more time consuming and the EM algorithm would require customized implementation in case of any changes in the model specifications.

The least squares method can be easily extended to multiple intercrosses among inbred lines or multiple families in outbred populations. However, its residuals are not truly *i.i.d.* $\mathcal{N}(0, \sigma_e^2)$ as discussed by XU and ATCHLEY [1995]. Instead, the residual variances are heterogeneous and inflated by the amount of the conditional variance of unobserved QTL genotypes given the observed marker information. XU [1998] used an iteratively re-weighted least squares (IRWLS) algorithm to account for this bias

in residuals but found little difference between least squares and IRWLS.

One noteworthy drawback of RIM concerning the outbred half-sib cattle populations is that it does not account for the maternally inherited QTL alleles, which causes residuals to have a mixture distribution rather than a single normal distribution [HOESCHELE, 2007]. In their applications for QTL mapping in outbred populations, neither IM and RIM separates polygenic effects out of the residual term, either.

Multiple QTL models

The Lander-Botstein IM is a single QTL model thus incapable of analyzing more than one QTL simultaneously in the same model, let alone the desired inference of interactions between QTL (*a.k.a.*, epistasis, referring to two-way interaction in this dissertation unless otherwise specified). In late 1980s and early 1990s, a number of proposals [JANSEN, 1993; ZENG, 1993, 1994] were raised to extend IM to a multiple QTL modeling device by combining IM with multiple regression in which markers or QTL not in the interested interval can be incorporated into the model to control the genetic variation and obtain better estimate of residual variations. Jansen’s method is known as MQM (multiple-quantitative trait loci mapping) and Zeng’s method is well known as composite interval mapping (CIM) in the literature.

CIM is still not a *bona fide* multiple QTL model because the covariates have to coincide with markers. KAO *et al.* [1999] presented in detail the multiple interval mapping (MIM) method that allows multiple QTL to be simultaneously modeled and these QTL can reside anywhere within the marker coverage region. One further advance in methodology from CIM to MIM is that MIM makes modeling of epistasis possible. There are continuing debates about whether an epistasis should be included into the model if neither or one of the interacting QTL’s main effect has not been included in the model, but we will not dwell much on the topic of epistasis modeling in this literature review.

It takes a whole set of notations, model specifications and description of statistical procedures of model selection and evaluation to present the MIM framework and we put off this task until Chapter 3 where we will present our adaptation and extension

of MIM for QTL mapping in outbred half-sib cattle populations.

1.3.2 Linear random effects model

Polygenic effects

Now we get to discuss QTL mapping methods that involves one or several random effects other than the residuals. One terminology often seen in the literature is polygenic effects – small effects by a large number of loci [FERNANDO *et al.*, 1994]. Although the genetic variance can also be estimated in crosses between inbred lines [BREM and KRUGLYAK, 2005], each genetic contribution (*a.k.a.*, QTL/genes) must be modeled explicitly for evaluation of the trait heritability. Reasons are that all the pairwise relationships among individuals in a cross between inbred lines are identical, thus the kinship coefficients among all pairs are also identical, which results in no power to estimate the polygenic component as a random effects. However, in cases of multiple intercrosses among multiple inbred lines, polygenic effects can be modeled under a variance components model framework [CREPIEUX *et al.*, 2004]. Even so, cautions are still needed to confine the interpretation of analysis results to only the inbred populations founded by the same parental lines and to the extent and pattern of IBD between them [SIEBERTS and SCHADT, 2007].

In outbred populations in which parents are of divergent background, especially in outbred populations of multiple pedigrees and/or multiple generations, polygenic effects needs to be taken into the model for proper estimation of the QTL heritability. Polygenic effects (denoted as \mathbf{u}) are generally assumed to be a random effect as $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{A})$ where the correlation matrix \mathbf{A} is twice of the kinship coefficient matrix based on pedigree relationship only. The kinship coefficient is zero for unrelated individuals and half for self or monozygotic twins. Table 1.3.2 summarizes twice of the kinship coefficients for some commonly encountered relative pairs and such relationships may become very tedious to trace recursively.

Table 1.2: Kinship coefficients for simple relative relationships.

Relationship	Twice of the Kinship Coefficient
Self or Monozygotic twin	1
Parent, offspring or full sibs	1/2
Half sibs, grand parent or grand child	1/4
Cousin	1/8

Fixed QTL effects with random polygenic effects

Inclusion of the random polygenic effects does not prevent QTL being treated as fixed effects. Considering a single QTL and polygenic background, a general form of the likelihood can be written as:

$$L(\mathbf{y}|\mathbf{M}) = \sum_{\mathbf{G}} \Pr(\mathbf{G}|\mathbf{M}) f(\mathbf{y}|\mathbf{G}) \quad (1.3.3)$$

with

$$\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V}) \text{ and } \mathbf{V} = \mathbf{Z}_u \mathbf{A} \mathbf{Z}_u' \sigma_u^2 + \mathbf{I} \sigma_e^2.$$

where

- \mathbf{y} is the trait phenotypes;
- \mathbf{M} is the marker information;
- \mathbf{G} is the vector of QTL genotypes (not necessarily assumed to be bi-allelic);
- \mathbf{u} is the vector of polygenic effects as defined above;
- \mathbf{Z}_u is an incidence matrix relating phenotypes in \mathbf{y} to individual's additive polygenic effect;
- σ_e^2 is variance of residuals that are assumed to be *i.i.d.*.

Difficulty exists in the summation over \mathbf{G} and the evaluation of the multivariate normal density $f(\mathbf{y}|\mathbf{G})$ because \mathbf{V} is not diagonal any more. For simple pedigrees, this

likelihood can be evaluated using numerical integration [MORTON and MACLEAN, 1974; KNOTT *et al.*, 1991]. Approximations to this likelihood or its variants have also been proposed such as modal estimation in HOESCHELE [1988]. Monte Carlo Markov Chain (MCMC) algorithms for sampling genotypes and/or polygenic effects can also evaluate this likelihood, especially in complex pedigrees (see discussions of MCMC's application in linkage analysis by HOESCHELE [2007] and THOMPSON [2007]).

Random QTL effects models

In this part, we discuss QTL mapping via variance components (VC) model in which QTL effects are treated as random and the variance components due to the QTL effect and polygenic effects are estimated usually by the restricted maximum likelihood (REML) method. In this respect, QTL effects are expressed in terms of a proportion of phenotypic variance explained by the QTL, *a.k.a.* the QTL heritability. Since the VC method is robust and generally applicable for a wide variety of population structures such as many or a few large or small families from simple or complex pedigrees, etc., as well as its lax assumption of the number and frequencies of QTL alleles, VC model for QTL mapping were developed independently by both the human genetics [ALMASY and BLANGERO, 1998] and animal breeding [GRIGNOLA *et al.*, 1996; GEORGE *et al.*, 2000] communities.

Consider a single QTL variance component model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_q\mathbf{q} + \mathbf{Z}_u\mathbf{u} + \mathbf{e} \quad (1.3.4)$$

where

- \mathbf{q} is a vector of additive QTL effects at a locus of interest;
- \mathbf{Z}_q is an incidence matrix relating individual's trait phenotype with random QTL effects vector;
- Other notations are defined in the same way as above.

The covariances between additive QTL effects and additive polygenic effects are

assumed to be zero based on the following arguments: (1) one QTL is fitted per marker interval; (2) the polygenic effect \mathbf{u} is the sum of a large number of tiny effects at all loci that are neither modeled as QTL nor linked with the modeled QTL; and (3) all QTL in the population are assumed to be in linkage equilibrium. Therefore, the variance components for model (1.3.4) can be written as $Var(\mathbf{y}) = \mathbf{V} = \sigma_q^2 \mathbf{Z}_q \mathbf{G} \mathbf{Z}_q' + \sigma_u^2 \mathbf{Z}_u \mathbf{A} \mathbf{Z}_u' + \sigma_e^2 \mathbf{I}$, where \mathbf{G} is the IBD coefficients matrix estimated from observed marker data and \mathbf{A} is twice of the kinship matrix.

REML is the preferred method to obtain a solution to this model, since the variance components are unknown and need to be estimated in order to maximize the likelihood. The testing hypothesis of QTL variance component can be formulated as $H_0 : \sigma_q^2 = 0$ vs. $H_1 : \sigma_q^2 > 0$. A likelihood ratio test can be constructed between the full model as specified in equation (1.3.4) and the reduced model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_u \mathbf{u} + \mathbf{e}$. Such an LRT on variance components leads to one further statistical topic, *i.e.*, the test statistic under the null hypothesis (no QTL at the test location) follows a 1:1 mixture distribution of χ_0^2 and χ_1^2 [SELF and LIANG, 1987].

In practice, however, the location of QTL is unknown and genome scan for mapping QTL is conducted. Though the point-wise null distribution of the LRT statistic on a QTL variance component is well defined as above, the distribution of the chromosome-wide or genome-wide maximum of the LRT statistics under the null hypothesis that there is no QTL on the chromosome (genome) is analytically unknown. Recall that in the Lander-Botstein interval mapping framework, genome-wide significance threshold can be obtained by permutations [CHURCHILL and DOERGE, 1994; DOERGE and CHURCHILL, 1996]; however, such strategy cannot readily be employed in general pedigrees, due to the need to preserve pedigree relationship thus the polygenic effects while permuting the association between trait phenotypes and marker (thus QTL) genotypes. Nevertheless, GEORGE *et al.* [2000] suggested that the chromosome-wide significance threshold can be approximated by a χ_1^2 distribution.

Variance component method for QTL mapping can also be extended to multiple additive QTL effects by involving much more complicated correlation structures between QTL and markers [ALMASY and BLANGERO, 1998]. Such a procedure of forward search for QTL has been implemented in the software package SOLAR

(<http://solar.sfbrogenetics.org/download.html>), in which the model selection criteria for adding a second (or third, etc.) QTL can be arbitrarily set and is generally recommended to be $\text{LOD} = 1.9$.

1.3.3 Bayesian QTL mapping

There are many variants of Bayesian QTL modeling in terms of single versus multiple traits, choices of priors, fixed or random QTL effects, and the core computational implementation of the MCMC algorithms such as Metropolis algorithm, Metropolis-Hastings algorithm, Gibbs sampler or reversible jump MCMC [SATAGOPAN and YANDELL, 1996; WAAGEPETERSEN and SORENSEN, 2001; YI *et al.*, 2005], etc.

Assume a simple scenario: a single QTL is located at position d in a marker interval (assuming the order as MQN) in an inbred F_2 population and the two flanking markers (marker genotypes and locations) are known. Denote \mathcal{Q} the unknown QTL genotype of all individuals and \mathcal{Q}_i the unknown QTL genotype of individual i . For quantitative traits, it is usually assumed that the trait phenotype of individual i takes the conditional normal distribution of the following form:

$$y_i | \mathcal{Q}_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma^2), & \text{if } \mathcal{Q} = QQ; \\ \mathcal{N}(\mu_2, \sigma^2), & \text{if } \mathcal{Q} = Qq. \\ \mathcal{N}(\mu_3, \sigma^2), & \text{if } \mathcal{Q} = qq. \end{cases} \quad (1.3.5)$$

where μ_1 , μ_2 and μ_3 are the mean phenotypic value of individuals with QTL genotypes QQ , Qq and qq , respectively, and the residual variance σ^2 . It is easy to see that half of the mean of μ_1 and μ_3 is the QTL additive effect (α) and the difference between μ_2 and the mean of μ_1 and μ_3 is the QTL dominance effect (δ), that is, $\alpha = (\mu_1 + \mu_3)/2$ and $\delta = \mu_2 - (\mu_1 + \mu_3)/2$. The parameters of the Bayesian model are thus (\mathcal{Q}, r_d) and $\theta = (\mu_1, \mu_2, \mu_3, \sigma^2)'$, where r_d is defined as the recombination fraction between the flanking marker M to the left of QTL and the QTL location d .

Choice of priors can be a topic of its own and we mention the commonly used conjugate priors as the following: the location of QTL is of Beta distribution (uniform distribution can be considered as a special case of Beta distribution), *i.e.*,

$\eta \sim Be(a, b)$, $0 < \eta < 1$, such that $r_d = \eta r_{MN}$, with a and b defining the shape of the distribution; the three mean phenotypic values μ_1 , μ_2 and μ_3 follows a zero-mean normal prior, say $\mathcal{N}(0, \sigma_0^2)$ where σ_0^2 are usually large so as to allow for large QTL effects and to be not-so-informative; the residual variance σ^2 is assumed to be from an inverse Gamma prior, in shorthand as $\sigma^2 \propto 1/\sigma^2$. The *a priori* independence between (\mathcal{Q}, r_d) and $\boldsymbol{\theta}$ is also assumed.

The posterior distribution of $(\mathcal{Q}, r_d, \boldsymbol{\theta})$ is given by

$$\begin{aligned}
 p(\mathcal{Q}, r_d, \boldsymbol{\theta} | \mathbf{y}, \mathbf{M}) &\propto p(\mathcal{Q}, r_d) p(\boldsymbol{\theta}) p(\mathbf{y}, \mathbf{M} | \mathcal{Q}, r_d, \boldsymbol{\theta}) \\
 &\propto p(\mathcal{Q}, r_d) p(\boldsymbol{\theta}) p(\mathbf{M} | \mathcal{Q}, r_d, \boldsymbol{\theta}) p(\mathbf{y} | \mathcal{Q}, r_d, \boldsymbol{\theta}, \mathbf{M}) \\
 &\propto p(r_d) p(\mu_1) p(\mu_2) p(\mu_3) p(\sigma^2) \Pr(\mathcal{Q} | r_d, \mathbf{M}) p(\mathbf{y} | \mathcal{Q}, \boldsymbol{\theta}) \\
 &\propto p(r_d) p(\mu_1) p(\mu_2) p(\mu_3) p(\sigma^2) \prod_{i=1}^n \Pr(\mathcal{Q}_i | r_d, \mathbf{M}_i) \prod_{i=1}^n p(y_i | \mathcal{Q}_i, \boldsymbol{\theta})
 \end{aligned} \tag{1.3.6}$$

where the last step involves conditional independence such that QTL genotype can be drawn from each individual at a time, which cannot be generalized to outbred populations.

According to equation 1.3.6, the marginal posterior distributions of $\boldsymbol{\theta}$, namely $(\mu_1, \mu_2, \mu_3, \sigma^2)$, can be readily derived, due to the above choice of conjugate priors. Their posterior distributions will follow normal distribution (for the mean parameters μ .) and inverse Gamma distribution (for residual variance parameter σ^2), respectively, with the posterior distributional parameters some functions of prior parameters and MLEs from the data.

The marginal posterior distribution of QTL genotype takes the following form:

$$\Pr(\mathcal{Q} | \cdot, data) \propto \prod_{i=1}^n \Pr(\mathcal{Q}_i | r_d, \mathbf{M}_i) p(y_i | \mathcal{Q}_i, \boldsymbol{\theta}). \tag{1.3.7}$$

Therefore, the posterior sampling of QTL genotypes can be drawn separately for each

individual, with the following formulae:

$$\left\{ \begin{array}{lcl} \Pr(\mathcal{Q}_i = QQ|., data) & = & \frac{\Pr(\mathcal{Q}_i = QQ|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = QQ, \boldsymbol{\theta})}{\sum \Pr(\mathcal{Q}_i = \omega|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = \omega, \boldsymbol{\theta})} \\ \Pr(\mathcal{Q}_i = Qq|., data) & = & \frac{\Pr(\mathcal{Q}_i = Qq|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = Qq, \boldsymbol{\theta})}{\sum \Pr(\mathcal{Q}_i = \omega|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = \omega, \boldsymbol{\theta})} \\ \Pr(\mathcal{Q}_i = qq|., data) & = & \frac{\Pr(\mathcal{Q}_i = qq|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = qq, \boldsymbol{\theta})}{\sum \Pr(\mathcal{Q}_i = \omega|r_d, \mathbf{M}_i)p(\mathbf{y}_i|\mathcal{Q}_i = \omega, \boldsymbol{\theta})} \end{array} \right. \quad (1.3.8)$$

where $\omega = QQ, Qq$ or qq .

The posterior distribution of the relative QTL location, r_d , takes the following form:

$$\Pr(r_d|., data) \propto p(r_d) \prod_{i=1}^n \Pr(\mathcal{Q}_i|r_d, \mathbf{M}_i). \quad (1.3.9)$$

Here comes a problem that the posterior distribution of r_d does not have a standard form and Metropolis-Hastings (MH) algorithm [METROPOLIS *et al.*, 1953] can be used to draw samples r_d such that the acceptance probability is:

$$\alpha(r_d^*, r_d) = \begin{cases} \min \left[\frac{p(r_d^*|.data)u(r_d|r_d^*)}{p(r_d|.data)u(r_d^*|r_d)}, 1 \right], & \text{if } p(r_d|., data) > 0; \\ 1, & \text{otherwise.} \end{cases} \quad (1.3.10)$$

where r_d^* denotes a candidate value generated by the candidate generating density $u(r_d^*|r_d)$.

The MCMC approach described above, once implemented, generates Monte Carlo samples from the joint posterior distribution as specified in equation (1.3.6), from which posterior inferences can be made. For the purpose of mapping QTL, the priority task is to draw inferences about whether there is QTL or not. Bayes factor (BF), a Bayesian counterpart to the classical likelihood ratio test in the frequentists' world, can be used to assess which of the models prevails: with or without QTL. Under the

model of no QTL (M_0), the posterior distribution is

$$p(\mu, \sigma_e^2 | \mathbf{y}, M_0) \propto p(\mu, \sigma_e^2 | M_0) p(\mathbf{y} | \mu, \sigma_e^2, M_0) \quad (1.3.11)$$

The Bayes factor of model with one QTL (M_1) relative to model M_0 is defined as follows:

$$BF = \frac{p(\mathbf{y} | M_1)}{p(\mathbf{y} | M_0)} = \frac{\int p(\mathcal{Q}, r_d, \boldsymbol{\theta} | \mathbf{M}, M_1) p(\mathbf{y} | \mathcal{Q}, \boldsymbol{\theta}, M_1) d\boldsymbol{\theta}}{\int p(\mu, \sigma_e^2 | M_0) p(\mathbf{y} | \mu, \sigma_e^2, M_0) d\mu d\sigma_e^2} \quad (1.3.12)$$

A Monte Carlo estimate of the Bayes factor above can be given by using the harmonic mean of the likelihood values from the Monte Carlo chain (of the length m), as suggested by NEWTON and RAFTERY [1994] in the equation (13) therein, as follows:

$$\widehat{BF} = \frac{\left\{ \frac{1}{m} \sum_{j=1}^m \left[p(\mathbf{y} | \mathcal{Q}^{(j)}, \boldsymbol{\theta}^{(j)}, M_1) \right]^{-1} \right\}^{-1}}{\left\{ \frac{1}{m} \sum_{j=1}^m \left[p(\mathbf{y} | \mu^{(j)}, (\sigma_e^2)^{(j)}, M_0) \right]^{-1} \right\}^{-1}} = \frac{\frac{1}{m} \sum_{j=1}^m \left[p(\mathbf{y} | \mu^{(j)}, (\sigma_e^2)^{(j)}, M_0) \right]^{-1}}{\frac{1}{m} \sum_{j=1}^m \left[p(\mathbf{y} | \mathcal{Q}^{(j)}, \boldsymbol{\theta}^{(j)}, M_1) \right]^{-1}} \quad (1.3.13)$$

where j denotes the j -th draw of the parameters from their respective posterior distributions, respectively.

Now we briefly discuss more complicated cases such as mapping multiple QTL and/or epistasis effects in outbred populations. The difficulty lies in all aspects: the number of QTL is not known; the inter-QTL interactions are not known; the polygenic background needs to be accounted; genotyping failures and uninformative markers reduce data information content; meioses between generations need to be tracked/inferred; the uni-directional QTL gene flow in the pedigree need to be taken into consideration when inferring QTL genotypes; etc. These complexities not only increase the hardship of retrieving the joint posterior distribution containing more and complex unknown parameters, but also involve dimensional changes in the parameter space when a QTL or epistasis term needs to be added or dropped, which was not properly addressed until the reversible jump MCMC [GREEN, 1995].

MCMC is the main weaponry to obtain sampling from the joint posterior distri-

bution and there are several types of samplers in use. In cases where the number of QTL (or precisely speaking, the dimension of QTL effects vector) is treated as fixed, thus the parameter space dimension remain the same through the Monte Carlo chain, the Metropolis algorithm [METROPOLIS *et al.*, 1953], Metropolis-Hastings algorithm [HASTINGS, 1970] and the Gibbs sampler algorithm [GEMAN and GEMAN, 2000] have all been used. The aforementioned Bayesian interval mapping of a single QTL in inbred F_2 design is one example where MH algorithm can be used, because the exact form of the posterior distribution of the QTL location r_d was not available. On the other hand, when the posterior distribution of an unknown quantity is explicit and easy to sample from, the Gibbs sampler can be used. This usually requires the use of conjugate priors. Gibbs sampler works in a sequential updating manner: one unknown is updated at a time given all other unknowns at the most current states and such updates iterate through the whole vector of unknowns, thus forming a Markov chain. After a long time of run, the desired posterior joint distribution is assumed to be achieved and samples from it are used for posterior parameter inferences. Diagnostics of the convergence of a MCMC remains a challenge and common practice is to evaluate/compare multiple independent chains and to have MCMC run for a large number of Monte Carlo iterations.

When the number of QTL itself is treated as an unknown, things gets more complicated since the dimensionality of the parameter space changes when a QTL effect term is added to or dropped from the model. One tentative approach is to use Bayes factor as a measure of comparison between models with alternative number of QTL effects, with each model obtained after Monte Carlo run(s). However, the methodological advance called reversible jump MCMC, proposed by GREEN [1995], however, makes it possible to sample from the desired posterior distribution even when the dimension of parameter space is not fixed and that the number of QTL effects itself can be sampled together with all other unknowns. There have been several reversible jump MCMC reported in literature such as SATAGOPAN and YANDELL [1996]; WAAGEPETERSEN and SORENSEN [2001]; YI *et al.* [2005], etc.

1.3.4 Nonparametric QTL mapping

All methods described above share one common assumption that the trait phenotype follows a normal distribution with respective residual variance assumptions, thus all belong to the parametric QTL mapping approaches. However, there are cases where phenotypes of interest are not normally distributed such as count, probability/ratio, survival time, categorical and/or ordinal traits. Various procedures have been proposed such as the generalized linear model approach [M. *et al.*, 1996], liability threshold model approach [FALCONER and MACKAY, 1996] and semi-parametric models [ZOU *et al.*, 2002]. Transformations of these non-normal traits to normal distributions so that parametric QTL mapping methods can be applied [WRIGHT, 1968] are also common practice, though there is no guarantee that appropriate transformation may be found or a transformation would work.

It is straightforward to construct a nonparametric test at genotyped markers. Take the inbred backcross as an example: the progeny can be classified as either homozygous or heterozygous at a marker, and the Wilcoxon rank sum test (Wilcoxon test), a nonparametric counterpart to the parametric two-sample t-test, can be performed to test for rank-based phenotype difference between the two groups. Statistical procedures for obtaining appropriate significance thresholds for the Wilcoxon test was readily available, too. KRUGLYAK and LANDER [1995] proposed a nonparametric QTL interval mapping method as an analog to the work of Lander-Botstein IM by using a variant of Wilcoxon rank sum test. COPPIETERS *et al.* [1998] adapted the above non-parametric rank-based test from inbred crosses to outbred half-sib grand daughter designs and the more complicated design variants [COPPIETERS *et al.*, 1999].

In the several simulation-based power analyses between parametric and nonparametric QTL interval mapping methods [KRUGLYAK and LANDER, 1995; COPPIETERS *et al.*, 1998, 1999], there was minimal loss of power when the traits are actually normally distributed. When the residuals were simulated as non-normal or heteroscedastic, the rank-based method usually outperforms the regression interval mapping (RIM) method. Above being said, one major disadvantage of the rank-based methods is the fact that they do not provide convenient estimates of QTL effects,

which restricts the usage of these methods to the tasks of detecting QTL only and is the reason that such nonparametric tests is not of wide use today.

1.4 Background and Motivations of Our Research

Cattle populations are usually based on paternal half-sibship. There are two major designs for cattle: the daughter design with large sets of paternal half-sisters being used and the grand daughter design (GDD) that consists of large sets of (paternal) half-brothers (Figure 1.1). The grand daughter design is more powerful than the daughter design in that it would only require one third number of genotyped animals to detect a QTL [WELLER *et al.*, 1990; GEORGES, 2007]. Because of the long life span of cattle as well as the long span from calf to adult, it will take decades of continuous efforts to maintain a daughter design and collect the needed trait phenotypes such as milk yield, let alone the more time-consuming GDD that spans three generations.

1.4.1 Genetic evaluation of dairy traits

Breeding value (BV) of an animal for a trait refers to the genetic merit of the animal. In animal model evaluation of dairy traits accomplished with a variance components model [WIGGANS and VANRADEN, 1989; VANRADEN and WIGGANS, 1991], BV is estimated based on the information of the animal itself and all known relationships this animal has. The true BVs of individuals, however, are generally unknown because (i) most traits are oligogenic or polygenic; (ii) for an individual, only a random half of its genes are transmitted to the offspring; (iii) the combination among these trait-affecting genes is random and the number of such combinations is large; (iv) the trait performance of the individual is also affected by environment. In statistics terminology, the estimated breeding value (EBV) is actually a BLUP (best linear unbiased prediction), but we will follow the tradition in the literature and use EBV anyway.

Half of EBV predicts the expected performance of its progeny relative to a popu-

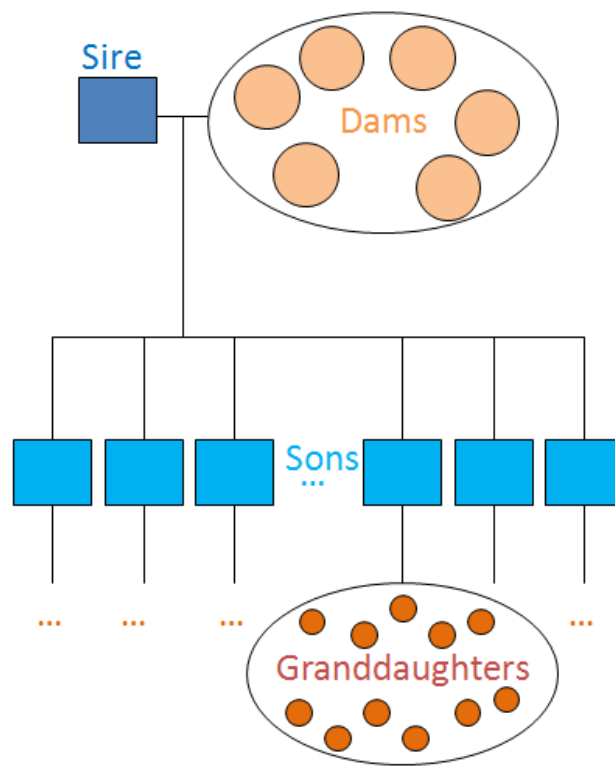


Figure 1.1: Diagram of grand daughter design in dairy cattle. Squares are bulls and circles are cows. Bulls (sire and sons) are genotyped but cows (dams and grand daughters) are not. The same bull mates with a number of dams and each gives birth to an offspring. Trait phenotypes of a bull are estimated from his own more than 50, sometimes more than 1000, daughters' trait performance.

lation mean under random mating, because for any offspring only half of the genes are transmitted from the animal and the other half are from the other parent at random from the population. This expected progeny performance as a difference from the population mean is called transmitting ability (TA) and the prediction of TA is the predicted transmitting ability (PTA).

Daughter (yield) deviation (DD or DYD) of a bull, on the other hand, is the evaluated weighted average of the daughters' yield deviations of its (10 or more) daughters that have been adjusted for the merit of their dams, respectively [VAN-RADEN and WIGGANS, 1991]. DD of a bull with granddaughters is estimated only from its daughters' information and the information of granddaughters and sons is excluded; therefore, DD offers an indication of a bull's daughters' performance without considering his parents or sons. DD also has the advantage of being unregressed and more independent, which makes it suitable for comparison of animals that are evaluated by different genetic evaluation models.

Reliability (REL), however, is not a measure of animals' genetic merit like EBV, PTA or DD; rather, it is a measure of the accuracy or degree of confidence in an evaluation. REL is defined as the squared correlation between an animal's true transmitting ability and PTA, and an equivalent definition of REL is the ratio between the variances of the animal's PTA and TA. Selection is ignored for computing REL in that correlations and variances are derived from unselected rather than selected population parameters. Like PTA, REL accounts for information from all relatives but only the terms for parents, self and progeny adjusted for mates are directly included in its estimation. REL is essentially a function of heritability of a trait (the proportion of the phenotypic variations that is due to genetic control) and the amount of information available for an animal. That information may come from the animal's own performance, from the performance of offspring, or from information for parents. As heritability and amount of information (such as a bull with more daughters) increase, REL also increases; the bigger REL is, the more accurate the evaluated PTA is. Therefore, when REL data is available, it is advised to incorporate it into the modeling to better account for the varying degree of accuracy in the estimated PTA or DD trait phenotypes.

1.4.2 Dairy cattle data for QTL mapping analysis

For the purpose of QTL mapping in dairy cattle populations, a genetic map of molecular markers is required. The Meat Animal Research Center (MARC) of United States Department of Agriculture (USDA) maintains an up-to-date bovine genome genetic map at <http://www.marc.usda.gov/genome/cattle/cattle.html> and we will refer to this map for marker locus information. The Haldane map function [HALDANE, 1919] is used under the assumption of no crossing-over interference.

In the grand daughter design [WELLER *et al.*, 1990], a grandsire and his sons are regarded as a paternal half-sib family. The grandsire is assumed to be segregating on one or a few QTL that influence the trait of interest. A large number of unrelated dams (cows) are mated with the grandsire and each giving birth to one son. In very rare cases should a dam giving births to two or more sons, one of the full-sibs is culled for the purpose of half-sib analysis. Each son is then mated with a large number of unrelated dams and his half-sib daughters are phenotyped. Through genetic evaluation using an animal model, the estimated breeding value (EBV) and yield deviation (YD) of daughters are obtained, then the predicted transmitting ability (PTA) and/or daughter (yield) deviation (DD or DYD) of the bulls are obtained, usually from the estimated genetic merits of more than 50 daughters through so-called progeny testing. These PTA or DD values are thus treated as the trait phenotypes of the bulls, although a bull itself won't give direct measures on traits such as milk production or pregnancy rate. Usually, only the grandsire and his sons are genotyped while cows' genotypes are unknown. The pedigree relationship among these bulls are also known and assumed to be error-free for the purpose of QTL mapping.

In a daughter design [WELLER *et al.*, 1990], similar scenario exists with the following distinctions: a paternal half-sib family consists of a sire and many of his daughters, each from an unrelated dam. The sire and his daughters are genotyped but dams are not. The EBV or YD of daughters are estimated through the animal model evaluation and the QTL is assumed to segregate in the sire.

1.4.3 Motivating data sets

The dairy cattle data set used in our research for method evaluations is from Dr. Melissa Ashwell’s lab (Department of Animal Science, North Carolina State University, Raleigh, NC, USA). This is a grand daughter design consisting of more than 900 animals that spans multiple generations [MUNCIE *et al.*, 2006]. In such a design, bulls (sires and sons) are genotyped while cows (dams and daughters) are not. A number of traits were retrieved from Animal Improvement Program Laboratory (AIPL) such as milk yields (MY), daughter pregnancy rate (DPR), productive life (PL), somatic cell score (SCS), calving ease (CE) and percent of difficult births (PDB). The phenotypes of these sons are the PTA or DD values estimated from each’s respective daughters that are usually more than 50 or even a thousand daughters by progeny testing. Therefore, these males’ estimated PTA or DD values are more accurate predictors of genetic potential than the cows’ values estimated from their own performances. Alongside the trait phenotypic PTA estimates, the estimated reliability (REL), a measure of accuracy, for each individual and each trait was also obtained. The genotype data contain 23 microsatellite markers located between 24.487cM (BMS2213) and 84.087cM (TGLA227) on the bovine chromosome 18 (BTA 18), among which some markers were severely uninformative or missing and five markers were only genotyped in a subset of offspring.

Preliminary QTL analysis of the data have been reported in MUNCIE *et al.* [2006] by using QTL Express (<http://qt1.cap.ed.ac.uk/>) [SEATON *et al.*, 2002] that implements the least squares regression method [HALEY and KNOTT, 1992]. Suggestive QTL were found for the traits of interest and improved inference of these QTL were sought by increased genotyped markers and improved methodology and software tools.

1.4.4 Outline of dissertation

In this dissertation, we present our work addressing various issues in the analysis flow of mapping QTL in outbred half-sib dairy cattle populations. The daughter and grand daughter designs of cattle populations and the large family sizes allow us to improve upon general formulations and make new developments in methodology as

well as the software Windows QTL Cartographer.

As a first step towards analyzing QTL in outbred half-sib populations, the haplotypes in the sires are needed for the purpose of QTL fixed effects modeling approach. Due to the artificial insemination technology, a sire could have dozens of or even hundreds of half-sib offspring. With genotypes of the sire and that of many of its offspring, it is of high probability that the marker linkage phase in the sire can be inferred with certainty. However, in even broader and more challenging scenarios where family size could be as small as 5; chromosomal segment could be as long as one or two Morgan; missing data could be as severe as 30%, etc., those currently available methods may not consistently perform well. In Chapter 2, we describe our new method of utilizing definite paternal alleles of genotyped markers and finding the most probable sire linkage phase. We demonstrate that our method is very robust and highly accurate and it has broad applicability towards small outbred families, long chromosomal segments with up to 30% missing data.

In Chapter 3, we report our extension of multiple interval mapping (MIM) method towards single half-sib family. MIM involves a maximum likelihood approach utilizing an Expectation-Maximization algorithm. In particular, we extend the MIM model to accommodate residual heteroscedasticity by incorporating reliability values estimated through genetic evaluations and demonstrate a better performance of QTL mapping. These new functionalities have been updated into our in-house QTL analysis software, Windows QTL Cartographer.

In Chapter 4, we describe our across-family QTL analysis under a mixed effects model framework. The rationale was to treat the QTL effects as random across multiple sires. However, we pose no restriction on the mean parameter of the random QTL effects, which allows alternative constructions of hypothesis testing that are shown to be better suited to capture more information from the data than a variance components model. We will also briefly introduce a simple approximated computation of IBD coefficients for half-sibships that we have been discussing all about and show the connection between our mixed-effects modeling and the variance components modeling.

Lastly, in Chapter 5, we summarize our work and main results in previous chapters

and discuss future direction(s) for QTL mapping, especially with the advent of the era of high density SNP markers. The much higher density of markers makes it possible to utilize the linkage disequilibrium (LD) information towards pinning down the trait-affecting marker alleles or haplotypes. Linkage analysis (LA) would still be valuable as it contains the “current” recombination events within the recorded pedigrees. On the other hand, LD offers insight into the historical recombination over generations prior to that of the pedigree founders. A joint utilization of LA and LD is a promising direction to pursue and there have been successful applications of such idea.

Chapter 2

Haplotyping in Outbred Half-sib Populations

2.1 Introduction

A haplotype refers to a particular set of alleles of closely linked loci. Alleles that are closely linked tend to be transmitted together from parent to offspring, therefore it would be of great advantage to be able to determine the haplotype configuration for an array of genetic analyses such as gene mapping, marker-assisted breeding, genetic counseling (concerning probability and/or diagnosis of hereditary diseases), etc. Current genotyping technology cannot offer experimentally-determined ordered marker genotypes (a.k.a., molecular haplotyping) at a high throughput pace and in a cost-effective way [JUDSON and STEPHENS, 2001]. It would help reduce the ambiguity in phasing if close relatives are also genotyped, however, the haplotyping could still become problematic when the number of loci is only moderately large. Missing values in genotype data also pose additional challenge for haplotyping. Thus, *in silico* haplotyping becomes an important computation task.

In outbreeding crosses between two diploidy individuals, the two parents are of divergent genetic background and there is no certainty what genotypes of a locus are in either or both parents. Denote Q as the locus and use subscript $1, 2, \dots$, to indicate different allele variants of co-dominant microsatellite (or SNP) markers, then there

could be up to four (two for SNP markers) different allele variants observed in two parents. The varying number of different allele variants in the two parents at one versus another locus leads to varying extent of informativeness along a chromosome in the parents. Assuming no missing data and no genotyping errors, the pair of parental genotypes Q_1Q_2 and Q_1Q_2 lead to unresolvable ambiguity in deducing parental origin of alleles in their offspring at this locus *per se*. Furthermore, if one or both parents are not genotyped at the locus, as is commonly practised in cattle half-sib populations, difficulty increases greatly in tracing the parental origin of offspring alleles. Additionally, the error-prone PCR-based microsatellite marker genotyping technology and genotyping failures due to bad DNA sample quality aggravate the extent of ambiguity in deducing parental origins of offspring alleles.

Cattle populations are usually based on half-sibship. There are two major designs for cattle: the daughter design with large sets of paternal half-sisters being used and the grand daughter design (GDD) that consists of large sets of (paternal) half-brothers. In usual practice, the sires and offspring (daughters in daughter design and sons in GDD) are genotyped but not the dams, which causes much uncertainty towards the task of haplotyping. The increasing marker density as well as large family sizes, which are easily seen in cattle populations due to extensive use of artificial insemination (AI), also present great challenges for *in silico* haplotyping because the space of all compatible haplotype configurations can be simply too large for exhaustive search [ELSTON and STEWART, 1971; LANDER and GREEN, 1987; WEEKS *et al.*, 1995; GEORGES *et al.*, 1995; KRUGLYAK *et al.*, 1996; DU *et al.*, 1998; GAO *et al.*, 2004]. On the other hand, rule-based haplotyping methods [WIJSMAN, 1987; KNOTT *et al.*, 1996; O’CONNELL, 2000; TAPADAR *et al.*, 2000; QIAN and BECKMANN, 2002; LI and JIANG, 2003, 2004; ZHANG *et al.*, 2005] do not fully use the inter-marker distance information and cannot handle cases in which recombination events during gametogenesis is non-negligible.

In the following, we describe in detail a likelihood-based method we propose for the task of inferring the haplotypes in sires based on the observed genotype data of the sire and his offspring within a paternal half-sib family. This task is one of the first encountered towards a comprehensive analysis flow of mapping quantitative trait

loci (QTL) that are segregating in the sire bull. Through extensive simulation, we demonstrate that our method is robust, fast, and most importantly, highly accurate in an array of challenging scenarios. Lastly, we discuss prospective extensions to our method for general problem settings and improved speed performance for a large number of markers.

2.2 Review of Selected Haplotyping Methods

Before we introduce our new method, we would like to introduce in detail several representative methods for inferring the most probable haplotypes in parents, some of which will be included into our method evaluation and comparison. For the simplicity of discussion and to avoid confusion, we would like to make a distinction between haplotype and gamete. They both refer to a particular set of alleles from closely linked loci. However, haplotypes refer to the phased genotypes in an individual; whereas a gamete produced by that individual refers to a particular haploid chromosome (or chromosomal segment under the consideration) derived from either or both of the two haplotypes in this individual, depending on whether crossing-over takes place during gametogenesis or not. Haplotypes in offspring can also be called respective parental gametes, but parental haplotypes may not be identical to one or any observed parental gametes, which is exactly why we would need to devise methods and develop tools to reconstruct haplotypes in accordance with observed genotype data.

2.2.1 Knott et al. (1996) method

Knott-Haley method was described in KNOTT *et al.* [1996] and has been implemented in their web-interfaced software package QTL Express (<http://qtl.cap.ed.ac.uk/>) [SEATON *et al.*, 2002]. It was a quick and simple approach to minimize the total number of recombination during parental gametogenesis for every consecutive pair of markers. For a half-sib sire family, it first identifies all the markers that are heterozygous in the sire. Homozygous markers in the sire do not segregate in this sire family thus is uninformative. They are omitted from further consideration. Then

every pair of adjacent informative markers in the sire are examined in the progeny. A two-marker linkage phase that minimizes the number of recombination is picked. If both phases are equally likely, one is selected at random.

Clearly, this method does not use all the information available from the half-sib progeny, namely the missing/uninformative genotypes in offspring. However, it runs fast and performs reasonably well for parents with large number of offspring. Another advantage of this haplotyping method is that it does not require a most accurate genetic map known *a priori*, as long as the ordering of markers are correct.

2.2.2 Georges et al. (1995) method

GEORGES *et al.* [1995] described a likelihood-based method for assigning marker linkage phase in sires based on the observed genotypes in the sire and his offspring in dairy cattle half-sib populations. This method was further used by COPPIETERS *et al.* [1998] and ZHANG *et al.* [1998], etc. We would refer to it as Georges method in this dissertation since its formulation was explicit for dairy cattle half-sib structures, although the general formulation for likelihood of pedigree data dates much earlier such as ELSTON and STEWART [1971], LANGE and BOEHNKE [1983] and BOEHNKE [1991]. Georges method computes the likelihood of the half-sib family data under all the compatible sire linkage phases and such likelihood is obtained by exhaustively enumerating all compatible sire gametes and the thus obliged dam gametes in each offspring. Markers are assumed to be in linkage equilibrium in the sire (so that each possible linkage phase is *a priori* considered equal) and dams (such that the probability of dam gametes can be calculated from allelic frequencies of maternal alleles), respectively. The marker linkage phase that gives the maximum likelihood is considered the most probable sire linkage phase.

Suppose a genetic marker map is known and there are x heterozygous markers in the sire thus 2^{x-1} possible sire haplotype configurations in total. The posterior probability of the sire haplotype configuration (H^s) given the data (one sire and n

half-sib offspring) can be written as follows:

$$p(H^s|data) \propto \pi(H^s) \prod_{j=1}^n \left[\sum_{g_j^s} [\Pr(g_j^s|H^s = H_i^s, \mathbf{M}_{obs}) \Pr(g_j^d|g_j^s, \mathbf{M}_{obs})] \right], \quad (2.2.1)$$

where

- H_i^s is the i -th possible sire haplotype configuration, $i = 1, 2, \dots, 2^{x-1}$;
- $\pi(H^s)$ is the prior probability density for H^s and is usually assumed to be equal for all configurations;
- \mathbf{M}_{obs} is the marker genotype observed on the sire and offspring;
- g_j^s is the sire gamete inherited by the j -th offspring given the sire phase H_i^s and bM ;
- g_j^d is the obliged dam gamete inherited by the j -th offspring given g_j^s and \mathbf{M}_{obs} ;
- $\Pr(g_j^s|H^s = H_i^s, \mathbf{M}_{obs})$ is the probability of sire gamete g_j^s in the j -th offspring given the sire haplotype H_i^s and \mathbf{M}_{obs} and is a function of recombination fractions between markers (Haldane map function);
- $\Pr(g_j^d|g_j^s, \mathbf{M}_{obs})$ is the probability of the obliged dam gamete inherited by the j -th offspring given the sire gamete g_j^s and \mathbf{M}_{obs} and is a function of the allelic frequencies of the maternal alleles at the markers.

From equation (2.2.1), it is imaginable that the total number of computation grows exponentially with the number of markers and polynomially with the number of half-sibs. Exhaustive enumerations of all possible sire haplotype configurations and all compatible sire gametes in each offspring with respect to a sire haplotype configuration quickly become computationally impossible. GEORGES *et al.* [1995] suggested 15 markers as a practical upper limit and this limit does not seem easy to be lifted in the near future. Another point worth noting is that Georges method assumes linkage equilibrium between all markers in the sire so that a flat prior is chosen for

all possible haplotype configurations in the sire. The calculation of $\Pr(g_j^d | g_j^s, \mathbf{M}_{obs})$ from the allelic frequencies of the maternal alleles at the markers, with dams not being genotyped, relies on the assumption that dams are randomly sampled so that the estimation of allele frequencies would be consistent and the product of maternal allelic frequencies would equal the probability of the dam gamete, *i.e.*, the linkage equilibrium is assumed between markers in dams, too. All these assumptions could deviate from the reality, especially on account of the fact that the effective population size for Holstein-Friesian dairy cattle is estimated to be in the hundreds at most [GEORGES, 2007].

2.2.3 Large scale haplotype phaser - LSPH (2006)

The large scale pedigree haplotyper (LSPH, available at <http://cowry.agri.huji.ac.il/>), is a heuristic rule-based haplotyping algorithm (as well as the software) developed by BARUCH *et al.* [2006]. This method aims at reconstruction of haplotypes in pedigrees with high density markers such as microsatellite or SNP markers for small chromosomal segments such that its assumption of no recombination within the segment would hold. LSPH method consists of several steps as outlined below:

1. Reconstruction of missing genotypes and detection of conflicts in genotypes according to Mendelian rules and pedigree relationships.
2. Deduce the definite parental origins whenever possible for every marker in the offspring.
3. Assign the progeny haplotypes to the parents from the offspring with the most complete haplotypes. Such haplotype is called the resolved haplotypes (RHap).
4. The RHap is assigned to each of the parent's progeny, where possible. Which haplotype to assign depends on the existence of a resolved marker that is heterozygous in the parental genotypes and has been resolved in the progeny.
5. Due to the no recombination assumption, even a single locus at which the parent origins of alleles can be deduced will identify which haplotype was inherited from

the parent to offspring.

6. The newly deduced haplotypes in offspring may be used to further complement RHap.
7. RHap gets improved each time an offspring is found to have a resolved marker that is not resolved in his parents.
8. Steps between 4 and 7 are iterated, *i.e.*, between from the offspring to parents and from parents to offspring, until all loci have been resolved in RHap.
9. RHap is used as a haplotype source for deducing all remaining unresolved loci in offspring.
10. Unrelated families are *in silico* haplotyped separately and related families are proceeded recursively with the offspring of the first family considered as the parent of a new family.

LSPH algorithm was implemented with a graphical user interface for Windows operating system only. The performance of LSPH was evaluated through extensive simulations in comparison with SimWalk2 and PedPhase2 methods, in which LSPH was shown to be faster and more robust with respect to the input requirements such as missing or discrepant genotypes than SimWalk2 [SOBEL and LANGE, 1996] and PedPhase2 [LI and JIANG, 2003, 2004]. LSPH was also shown to offer comparable, if not superior, performance in terms of percentage of correct allele determination with that of PedPhase2 [BARUCH *et al.*, 2006].

2.2.4 Windig-Meuwissen (2004) method

WINDIG and MEUWISSEN [2004] proposed a rapid iterative algorithm for inferring the most probable sire linkage phase given the offspring marker genotypes. This method is capable of handling missing values as well as non-informative markers in the offspring. The Windig-Meuwissen method has the following set up:

1. As an initial sire linkage phase setup, arbitrarily assign one allele of a marker into one haplotype and the other allele into the other haplotype; repeat this for every heterozygous marker in the sire.
2. In each offspring and at each marker, determine the parental origin (maternal, paternal or unknown) of each marker allele.
3. For every consecutive marker pair in the sire, find in each offspring the closest informative markers to both sides that contain this marker interval (such pair of informative markers may vary from one offspring to another) and determine the genetic distance, thus the recombination fraction (θ) according to a (Haldane) map function, between the pair of informative markers in each offspring.
4. Determine whether the phase of these informative markers is in agreement with the designated sire haplotype: for marker pairs in agreement, the probability of a phase change is θ and if not in agreement $1 - \theta$;
5. Compute for each marker interval in the sire the probability of a phase change using the following equation:

$$P_{\text{switch}} = \frac{\prod_{i=1}^y \theta_i \prod_{j=1}^x (1 - \theta_j)}{\prod_{i=1}^y \theta_i \prod_{j=1}^x (1 - \theta_j) + \prod_{i=1}^y (1 - \theta_i) \prod_{j=1}^x \theta_j} \quad (2.2.2)$$

where there are x offspring with a marker phase different from the designated parental haplotype and y offspring with marker phases that are in agreement. The recombination fractions (θ 's) between the two closest informative markers may vary among offspring.

6. Find the sire marker interval that corresponds to the maximum probability of switch: if the probability is larger than 50%, change the sire haplotype phase by one crossing-over in this marker interval in the sire.
7. Repeat (4) to (6) until probabilities of switch for all marker intervals are less than 0.5.

This method is reasonably fast (*i.e.*, suitable for running on a personal computer), capable of handling dense marker data as well as large family size such as tens or hundreds. According to the authors, there would be some difficult situations, such as single offspring family (where the non-recombination-needed haplotype is more probable) or that both parents have the identical heterozygous markers such that their heterozygous offspring does not provide information on the parental haplotype. Although this method does not guarantee to give the correct sire haplotype, all the methods would likely have difficulties finding the correct haplotype in those difficult situations as well.

DRUET *et al.* [2008] reported their efforts in developing a method based on Windig-Meuwissen’s method for inferring sire haplotypes in dairy cattle and then using QIAN and BECKMANN [2002] rule-based method for inferring parental haplotypes in offspring given the phased sire haplotypes. However, DRUET *et al.* [2008] did not provide a complete method for handling cases where a single crossing-over is to take place between two informative markers while there are multiple non-informative markers situated in-between (personal communications). In such cases, choice of different locations for crossing-over would lead to different, but all possible, parental haplotypes.

2.3 Our Method for Inferring Sire Haplotypes

When considering QTL linkage mapping in dairy cattle half-sib populations, with respect to the marker linkage phase inference, we are only concerned with the sire haplotype reconstruction. The typical large number of paternal offspring, due to extensive use of artificial insemination, make the task possibly with high accuracy. However, as the number of markers increases, the length of marker coverage region increases (thus recombination becomes non-negligible) and the number of genotyped offspring decreases, such task becomes more challenging to be accurately performed. Missing and uninformative genotype data also tampers with the speed and accuracy performances.

We propose a new method that was aimed at resolving the above issues and offer fast speed, robustness against recombination and missing data as well as high

level of accuracy. It is a likelihood-based method related to Georges method but gets rid of the overly enumerations by Georges method due to missing/uninformative markers in an offspring. For every offspring, our method will identify paternal alleles at the informative markers only. Then given a designated sire linkage phase, we could calculate the probability of the observed sire gamete based on Mendelian laws and the recombination fractions between such informative markers. The most likely sire linkage phase is the one with the maximum *a posteriori* probability (MAP).

2.3.1 Reconstruction of sire haplotypes

Let \mathbf{M}_{obs} denote the marker map information such as marker positions and recombination fractions between markers (Haldane map function) as well as the observed marker genotypes in the sire and his offspring. Suppose a sire who has n half-sib offspring and there are x segregating markers in the sire. Then there are 2^{x-1} possible haplotype configurations in the sire. The posterior probability of the sire haplotype configuration (H^s) given the data (one sire and n half-sib offspring) can be written as follows:

$$p(H^s|data) \propto \pi(H^s) \prod_{j=1}^n \Pr(g_j^s | H^s = H_i^s, \mathbf{M}_{obs}) \quad (2.3.1)$$

where

- H_i^s is the i -th possible sire haplotype configuration, $i = 1, 2, \dots, 2^{x-1}$;
- $\pi(H^s)$ is the prior probability density for H^s and is usually assumed to be equal for all configurations;
- g_j^s is the definite sire gamete inherited by the j -th offspring given the sire phase H_i^s and \mathbf{M}_{obs} ;
- $\Pr(g_j^s | H^s = H_i^s, \mathbf{M}_{obs})$ is the probability of sire gamete g_j^s in the j -th offspring given the sire haplotype H_i^s and \mathbf{M}_{obs} and is a function of recombination fractions between markers.

Due to genotyping failures and non-informativeness where the parental origin of marker alleles cannot be unequivocally determined (e.g., the sire and the offspring have the identical heterozygous genotypes), the number and locations of such loci with definite paternal alleles may vary from one offspring to another. Our method identifies those markers with one allele of unequivocal paternal origin (informative markers) and skips any markers where the paternal allele cannot be determined clearly, and thus the sire gamete inherited by the j -th offspring, g_j^s , is deduced. Note g_j^s is definite given the observed genotypes in the offspring and the sire, no matter what the marker linkage phase in the sire is. The inherited sire gamete is then checked against the i -th sire haplotype configuration, H_i^s , to determine whether and where there would be recombination(s) during gametogenesis. In mathematic formulation, the probability of the sire gamete given the i -th sire haplotype configuration and the observed inherited sire gamete in the j -th offspring is calculated as follows:

$$P(g_j^s | H_i^s, \mathbf{M}_{obs}) = (1 - r_0) \cdot \left\{ \prod_{l=1}^m r_l^{\Delta_l} (1 - r_l)^{1 - \Delta_l} \right\} \cdot (1 - r_m) \quad (2.3.2)$$

where

- m is the number of fully informative markers in the j -th offspring;
- r_0 is the recombination fraction between the first marker and the first informative marker and r_m is the recombination fraction between the last informative marker and the last marker;
- r_l is the recombination fraction between the l -th and $(l + 1)$ -th informative markers;
- Δ_l is the indicator variable whether there would be a crossing-over during gametogenesis in the sire between the l -th and $(l + 1)$ -th informative markers with respect to this offspring j , using information from the i -th sire linkage phase and the deduced sire gamete.

Such computation varies from one offspring to another, depending on the observed genotypes. The formula (2.3.2) above also implicitly assumes that there is no recombi-

nation between the most outside informative marker and the most outside genotyped marker, if the most outside genotyped markers are not informative.

2.3.2 Inference procedures of most probable sire haplotypes

The procedure of our method for a paternal half-sib family is given below:

1. For each family (a sire and its progeny), determine segregating markers in the sire and enumerate all possible haplotype configurations.
2. For each offspring, determine the definite paternal allele at each marker. If a marker is not informative such that its paternal allele cannot be deduced unequivocally, this marker will be skipped.
3. Construct a sire gamete for the offspring containing those paternal alleles only and denote it as, g_j^s , the observed sire gamete for offspring j . See Figure 2.1 for a schematic diagram.
4. For a given sire haplotype configuration, compute the probability of g_j^s given the sire haplotype configuration and marker map information using the equation (2.3.2).
5. The likelihood of the haplotype configuration given the sire and offspring genotypes is proportional to the product of probability of individual sire gametes because of our assumption of independent gametogenesis.
6. Repeat steps (4) to (5) above for all possible haplotype configurations and find the posterior mode, *i.e.*, one haplotype configuration that maximizes the posterior probability (equation 2.3.1). The assumption of equal prior probability is usually used but other informative priors could also be invoked (see Future Work section in this chapter).

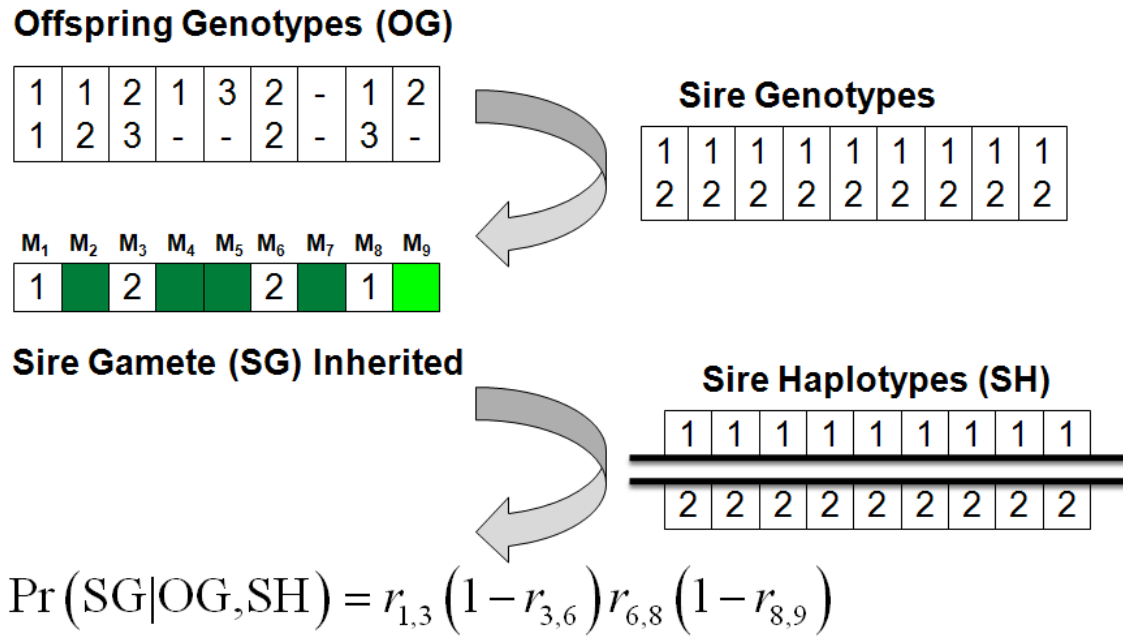


Figure 2.1: Deduction of sire gametes (SG) in the offspring based on offspring's genotypes (PG) and the sire haplotype (SH) configuration. $r_{x,y}$ represents the recombination fraction between the x -th and y -th markers, etc. Yellow-colored squares represent markers that are non-informative between a pair of fully informative markers whereas green-colored squares represent markers that are non-informative but outside the last informative markers in progeny.

2.4 Simulations

Our sire haplotyping method was implemented in C⁺⁺. To evaluate our method and compare it with other existing methods, we also implemented the Knott-Haley method [KNOTT *et al.*, 1996] and the Georges method [GEORGES *et al.*, 1995]. We further included the large scale pedigree haplotyper (LSPH) into our comparison. The Windig-Meuwissen method was not included because we tried in vain to obtain the Fortran program from neither author.

The accuracy, robustness and speed of our method were evaluated using data simulated from various scenarios with different lengths of chromosomal segments, number of progeny and the extent of missing values. The simulated chromosomal segment analyzed consisted of 11 evenly spaced markers. The number of progeny per sire varied from 5, 10, 20, 50 and 100 and the length of segment varied from 3cM, 10cM, 30cM, 60cM and 100cM. For each combination of family size and chromosomal length, ten independent replicates were simulated and each replicate contains 200 unrelated sire-sons families.

Genotypes were assumed known only for males. Markers were assumed to be poly-allelic with four alleles per marker such that the expected heterozygosity of 67% is comparable to what is observed in our experimental data (Table 3.4.1) as well as in reality with microsatellite markers in cattle populations [COPPIETERS *et al.*, 1999]. Population allelic frequencies were determined for each marker by simulation from a uniform distribution and re-scaled to sum to unity. Genotypes and haplotypes in sires were generated by random sampling of alleles on the basis of simulated population allelic frequencies. With respect to each offspring, the number of recombination events was simulated from a Poisson distribution with the parameter equal to the length of chromosome segment (in Morgans) and the locations of recombination, if any, were simulated from a uniform distribution. Either of the two sire gamete is then inherited by the offspring at equal probability. The dam gamete of each individual was determined by random sampling of alleles on the basis of the population allelic frequencies. In addition, different percentages (10%, 20% and 30%) of marker genotypes were randomly designated as missing.

The accuracy of sire haplotype reconstruction reflects the proportion of correctly reconstructed sire haplotypes out of the 200 independent founder sires within a simulated population. The speed of different methods was evaluated in terms of CPU time. Three methods (Knott-Haley, Georges and our method) were compared on a RedHat Linux cluster with dual Intel Xeon processors (2.8 - 3.2GHz) and 4GB memory per node. Analysis of data using LSPH was an exception because this program only runs on a Windows platform and its CPU time cannot be obtained. Instead, its calendar elapsed time (longer than the respective CPU time) was obtained from a desktop PC with a 2.8GHz Intel Pentium D CPU and 3.5GB of memory.

2.5 Results

Since chromosomal length does not seem to affect the speed performance (data not shown), we plotted the running time for haplotype inference versus different family sizes and extents of missing data (Figure 2.2). The greatly different scales in time of the y -axis clearly show that Knott-Haley method was the fastest and our method the second fastest. Georges method was the slowest because it needs to do many more enumerations and the calculations involved per enumeration are also more time-consuming. One interesting fact is that as the extent of missing data increases, Georges method takes longer time to finish while our method takes less time. The reason is, as more marker genotype data become missing, our method will skip those non-informative markers and have fewer computation tasks to do; whereas Georges method, as trying to enumerate all possible sire and dam gametes, will have to do more computations.

Figure 2.2 suggests that the extent of missing values does not affect the speed performance too much. Therefore, we averaged the running time for each method across different chromosomal lengths, extent of missing data, and simulation replicates (Table 2.5). The Knott-Haley method finished the sire haplotype reconstruction within a second in CPU time for a simulation replicate. Our method takes about 5 - 45 seconds in CPU time for the same amount of data and the Georges method takes

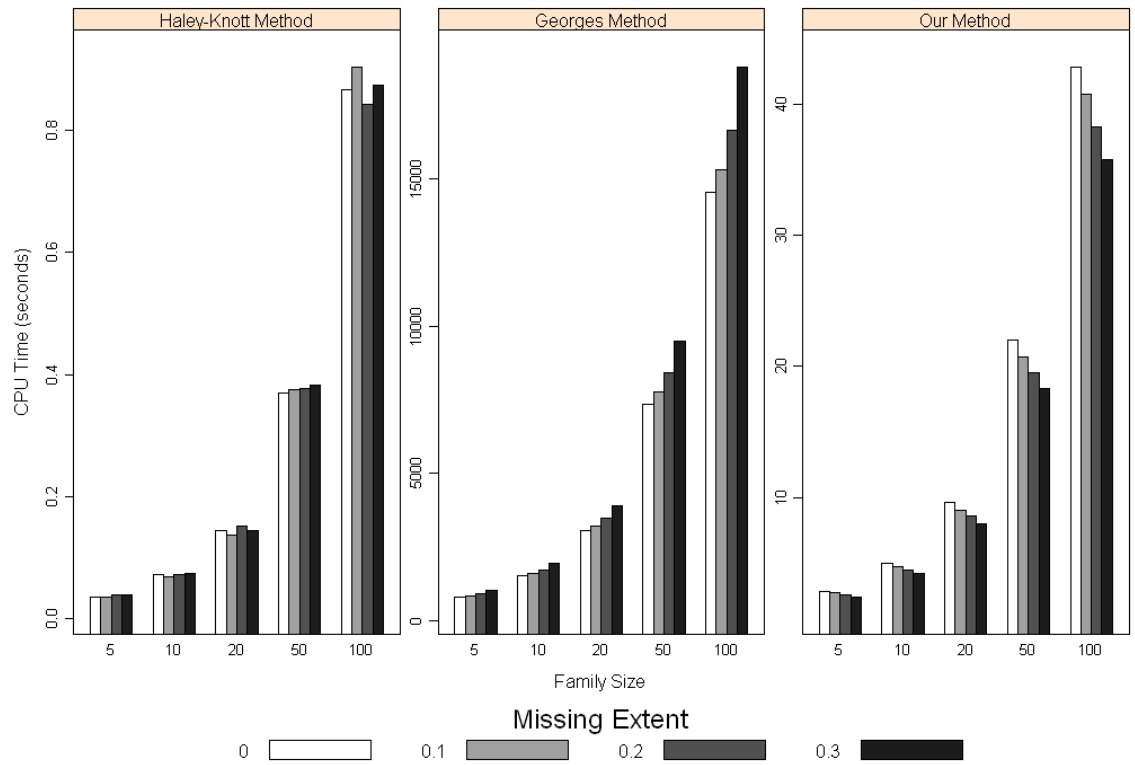


Figure 2.2: Speed performance (CPU time) comparison of haplotype reconstruction methods across different family sizes and extents of missing data.

Table 2.1: Speed performance comparison among sire haplotype reconstruction methods. CPU time for Knott-Haley, Georges and our methods was measured in seconds. The mean CPU time and standard errors were calculated across different chromosome lengths, various missing value extents and simulation replicates. The purple asterisks (*) denote entries for LSPH method that was the approximate calendar elapsed time in seconds.

Family Size	Haley-Knott	Georges	Our Method	LSPH*
5	0.037 (0.0004)	910 (9)	2.6 (0.019)	< 1*
10	0.072 (0.0006)	1701 (17)	4.6 (0.034)	< 1*
20	0.144 (0.0011)	3411 (32)	8.8 (0.087)	2*
50	0.376 (0.0032)	8261 (89)	20.1 (0.171)	4*
100	0.872 (0.0117)	16347 (174)	39.4 (0.438)	18*

tens of minutes or even hours. As a side note, it takes LSPH method between 1 and 18 seconds in calendar elapsed time (not the CPU time) to finish one simulation replicate, suggesting that LSPH method is faster than our method and Georges method.

The accuracy of reconstructed sire haplotypes is very important in method evaluation. We investigated a total of 100 cases comprising five different family sizes, five different lengths of chromosomal segments and four different extents of missing values in genotypes. Among all cases except six, our method achieves the greatest accuracy. The six cases in which Georges method gives the greatest accuracy all correspond to 100cM chromosome segment and a small family size of 5 or 10, (see the top row in Figure 2.3). By regarding each factorial combination as one Bernoulli trial and assigning a success if our method achieves the greatest accuracy among the four methods, we have 94 successes out of the 100 Bernoulli trials. We performed a binomial test in which the null hypothesis was the probability of our method achieving the greatest accuracy equaling 0.5 and the p -value was very small ($< 2.2 \times 10^{-16}$). In the contrast, similar binomial tests with respect to the other three methods were also conducted but none had a p -value smaller than 0.5. These test results suggest that our method is the best method in terms of accuracy.

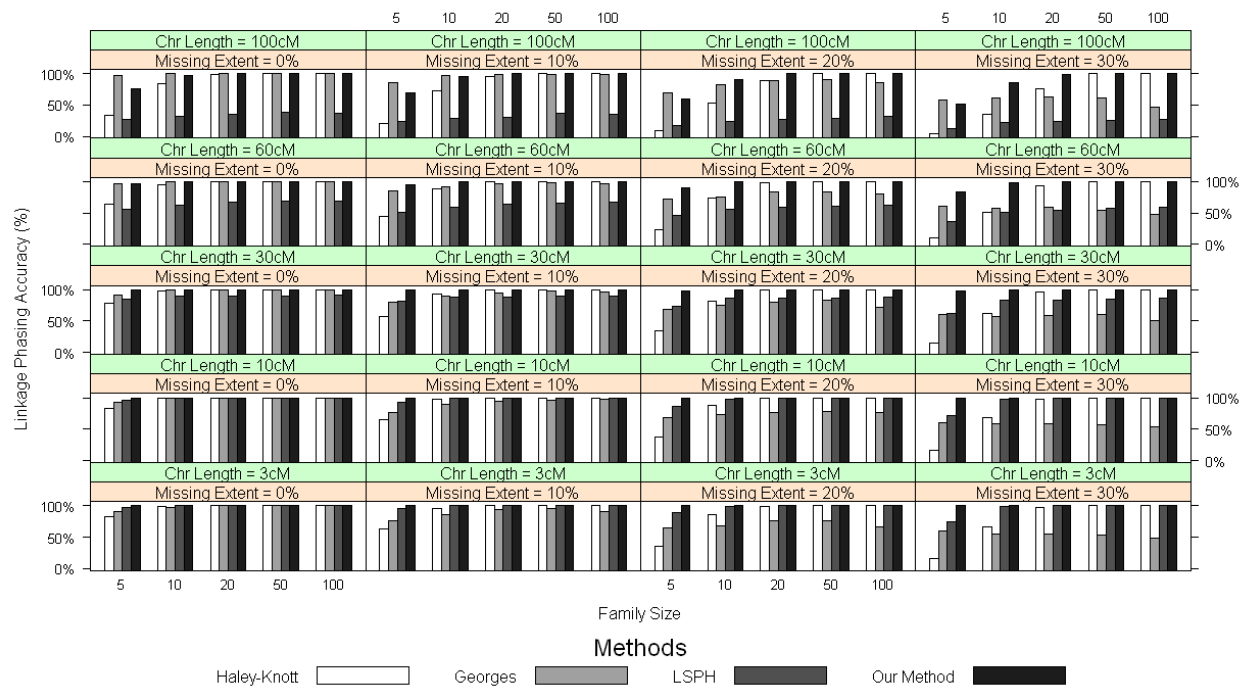


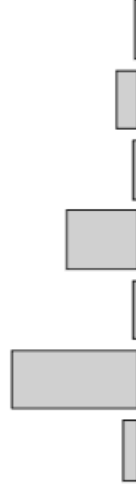
Figure 2.3: Accuracy of haplotype reconstruction versus family size, extent of missing data and length of chromosomal segment.

To better understand how well our method performs among the four methods as well as to illustrate various cases in which two or more methods tie for the greatest accuracy rate, we created a Venn diagram shown in Figure 2.4. Smiling faces in Figure 2.4 indicate that the corresponding methods are the greatest-accuracy-achieving methods and multiple smiling faces in the same row suggest ties. The column of counts (and the bar chart) to the left indicates the number of simulation settings in which a particular pattern of ties among the four methods is observed. For instance, in 9 out of 100 simulated cases, our method achieves the greatest accuracy along with Knott-Haley and Georges methods; whereas in 51 cases, our method was the sole most accurate method, etc. On another aspect, the counts in the bottom row of Figure 2.4 indicate the number of cases in which a method achieves the greatest (or equally greatest) accuracy. For example, our method achieves the greatest accuracy in 94 out of all the 100 simulated cases except the six cases in which Georges method gives the greatest accuracy; Knott-Haley method gives greatest accuracy in 41 cases which, however, are all ties with some other methods. In this respect, our method has clearly demonstrated its outstanding robustness and performance across a wide range of practical situations.

2.6 Discussions

There are also other methods of inferring linkage phase in the sire such as SimWalk2 [SOBEL and LANGE, 1996] that uses Markov chain Monte Carlo and simulated annealing algorithm, the rule-based methods by WIJSMAN [1987]; QIAN and BECKMANN [2002]; ZHANG *et al.* [2005], etc. However, since we have had LSPH included as one example of rule-based method assuming no recombination and that LSPH was developed with an aim at applications towards dairy cattle genetic marker data, we did not include other general-purpose rule-based methods into the method evaluations.

Among the four methods in our comparison, Knott-Haley method is the simplest and easiest to implement and the fastest method; Georges method is the most computationally expensive and is practically unusable when there are about a dozen of or



Counts	Our Method	Haley-Knott	Georges	LSPH
1	😊	😊	😊	😊
9	😊	😊	😊	
2	😊	😊		😊
29	😊	😊		
2	😊		😊	
51	😊			
6			😊	
Sub Total	94	41	18	3

Figure 2.4: Venn diagram of best-accuracy-achieving methods. Smiling faces stand for a method achieving the best accuracy rate and multiple smiling faces in the same row indicates tie in the accuracy performance among methods. The column of counts (and the bar plot to its left) gives the number of factorial combinations in which the corresponding patterns of smiling faces are observed. The subtotal row summarizes the total number of combinations in which a method gives the best accuracy. For instance, in 9 out of 100 simulated cases our method achieves the best accuracy along with Knott-Haley and Georges methods whereas in 51 cases our method was the sole most accurate method, etc. On another aspect, our method achieves the best accuracy in 94 out of 100 simulated cases except the six cases in which Georges method gives the best accuracy, etc.

more markers [GEORGES *et al.*, 1995]. Our proposed method has been demonstrated to be accurate, robust and fast across an array of different scenarios. Microsatellite marker data presents somewhat different challenges than SNP (single nucleotide polymorphism) markers in that the genotyping failure is more frequent and the recombination during meiosis is likely when the inter-marker distance is relatively large. Therefore, our method is advantageous in scenarios where recombination during meiosis is non-negligible.

Our method of haplotype inference is computationally fast. The more missing values, the faster our method because more markers would become non-informative and are skipped (Figure 2.1). On the other hand, more missing values would lead to lower accuracy in sire haplotype reconstruction due to the loss of information.

In terms of the accuracy of reconstructing sire haplotypes, our haplotyping method is the most robust against increasing recombination events (longer chromosomal segments), decreasing family sizes and increasing extent of missing data and attains the best accuracy among the four methods in most of the scenarios we considered (Figure 2.4).

The varying chromosome segment length, extent of missing genotypes and the number of offspring have different impacts on the other three methods, respectively. As the length of chromosomal segment increases (from bottom to top in Figure 2.3) such that recombination during spermatogenesis becomes non-negligible, LSPH method starts to fail and becomes unusable for 60cM or longer chromosomal segments, which is expected because LSPH was designed to work under the assumption of no recombination [BARUCH *et al.*, 2006]. In contrast, accuracy rate from Knott-Haley, Georges and our methods are all robust against recombination. On the other hand, as the extent of missing data increases (from left to right in Figure 2.3), Knott-Haley and Georges methods are all severely affected. LSPH method remains highly accurate as long as its assumption of no recombination is not severely violated. As the number of progeny increases (from left to right within each plot in Figure 2.3), the accuracy of most methods increases because the more progeny, the more information about the sire gametes and thus, the better inference of sire haplotypes. The non-increasing trend in accuracy of Georges method with respect to the increasing

family size, especially for short chromosomal segment, reflects the complex interaction among the varying settings in family size, recombination and missing values, the exhaustive enumeration of all possible sire gametes by Georges method, and the assumption of marker linkage equilibrium in dams that may not hold.

2.7 Future Work

There are many prospectives of extending our haplotyping methods, especially to improve its speed performance so that our method will be practically applicable towards hundreds of or even more markers on a chromosome. Another prospective would be to extend this method to outbred nuclear families in which haplotypes of both parents are reconstructed. A further extension is to develop a new module to infer most probable offspring haplotypes after the parental haplotypes having been inferred.

2.7.1 Speed performance improvement

Although our method is fast, it could become awkwardly time-consuming when say tens of hundreds of dense markers on a long chromosome need to be phased. One direction would be to effectively reduce the amount of computation by only considering those candidate haplotype configurations whose likelihood are top ranked, say the top 1,000 or 10,000 among all possible haplotype configurations. By this way, the computation time (for the top 1,000 or 10,000 haplotype configurations plus some overhead for picking these top-ranked configurations) are short and relative constant. The question then remains how to determine whether a candidate haplotype configuration would be top-ranked or not. Conversely, we need to be able to efficiently eliminate majority number of the all possible candidate haplotype configurations, which increases exponentially with the number of heterozygous markers in the sire, from likelihood calculation as detailed before.

We propose to use a robust and non-parametric index to quickly rank those haplotype configurations that in general correspond to larger likelihood. Such index is

a vector of integers with $x - 1$ elements, denoted as $\mathbf{t}_i = (t_1, t_2, \dots, t_k, \dots, t_{x-1})'$. The value $x - 1$ is the upper limit of the number of crossing-over, one per every consecutive marker interval. Each integer element, say t_k , is the count of offspring whose deduced definite sire gamete must have undergone k number of crossing-over according to a particular linkage phase H_i^s . In practice, the dimension of the index vector can be shrunk to only four by preserving the first three elements and the four element the sum of all counts corresponding to triple or higher order recombination.

For the i -th of 2^{x-1} sire haplotype configurations, such an index \mathbf{t}_i can be obtained. Intuitively, a candidate haplotype configuration with an index vector of much larger values for the first three elements, which correspond to the numbers of offspring who inherit their paternal gametes from the sire with zero, one or two crossing-over within the chromosomal segment being studied, would correspond to a larger likelihood than other configurations with index vectors having much smaller values for the zero and/or one recombination elements. Since such index vectors would be very fast to obtain, the speed for obtaining the top-ranked configurations would be polynomial with the number of heterozygous markers and the number of offspring. By this way, for a total number between 50 or 100 heterozygous markers on a chromosome such that there is on average one marker per $1 \sim 3$ cM or even denser, we can effectively reduce the number of likelihood computations from $2^{49} \approx 10^{15}$ or $2^{99} \approx 10^{30}$ to 1,000 or 10,000. What's more, such index vectors for candidate haplotype configurations can also be transformed into scalar prior weights to be used in equation (2.3.1).

2.7.2 Extension to mix sibship

Another prospective extension is to apply our method to full-sibs or mix-sibs families in which both parents or common parents are genotyped and need to be phased. Given a reasonably large family size (> 5 offspring), slight extent of missing data and a 100cM or shorter chromosomal segment, our method almost always gives 100% accuracy (Figure 2.3). The marker linkage phase in either parent can be inferred in the same way as previously described by ignoring the genotypes of the other parent. However, if both parents are genotyped, the deduction of parental gametes in an

offspring could contain alleles at more marker loci due to added information from the other parent. For example, suppose the genotypes of a locus Q are Q_1Q_2 in the sire, Q_1Q_3 in the dam and Q_1Q_2 in an offspring. The knowledge of the dam's genotypes make it trivial to assign Q_1 as the maternal allele and Q_2 as the paternal allele for this offspring. In this sense, the joint use of observed parental genotypes will improve the haplotyping.

2.7.3 Inference of offspring haplotypes

Although our method suffice our needs for QTL linkage mapping in half-sib populations, it would be meaningful to extend its functionality to also infer the haplotypes in offspring, after the parental haplotypes have been inferred. Common strategies include Monte Carlo methods [MEUWISSEN *et al.*, 2002; LEE *et al.*, 2005] to assimilate meioses in which the offspring haplotypes are assigned by a Gibbs sampler, various E-M algorithm-based methods [DING *et al.*, 2006], or some deterministic likelihood-based method [GAO and HOESCHELE, 2008], etc.

One straightforward way of assigning offspring haplotypes, given the offspring genotypes and phased parental haplotypes, could be deterministic assignment of offspring haplotypes by always choosing the mid-point as the recombination spot between two markers that require one crossing-over in between according to the definitely derived gametes in offspring and the known parental haplotypes. Double recombination or recombination of high order will always be ignored.

Chapter 3

Multiple Interval Mapping in Large Half-sib Families

3.1 Introduction

Methods for mapping genes for quantitative traits of economic importance in livestock such as milk production in dairy cattle need to account for the pedigree information over generations and marker heterozygosity because development and crossing of inbred lines is not feasible. Various approaches have been proposed with different assumptions on the number of QTL alleles and different methods for modeling QTL effects. When analyzing data from multiple families, there are two ways to analyze the data, either analyzing each single family separately (within-family analysis) or analyzing all families jointly (across-family analysis). The maximum likelihood (ML) and Bayesian methods usually assume QTL to be bi-allelic within a family as an approximation because it is difficult to ascertain the exact number of QTL alleles in outbred populations. The non-parametric methods proposed by COPPIETERS *et al.* [1998, 1999] also assume bi-allelic QTL, but this method has difficulty in estimating the desired QTL effects. However, selecting one or two large families may fail to detect a QTL unless this QTL happens to be segregating in these few families. The limitation of sample size for single family analysis also limits the statistical power of QTL detection. On the other hand, it is more reasonable to assume more than

two alleles at the QTL across multiple outbred populations, which makes methods such as the least squares (LS) [HALEY and KNOTT, 1992; MARTINEZ and CURNOW, 1992] and random effects models [AMOS, 1994; XU and ATCHLEY, 1995; GRIGNOLA *et al.*, 1996; ALMASY and BLANGERO, 1998; GEORGE *et al.*, 2000] more appropriate, because they make no specific assumptions about the number of QTL alleles. Random effects models also have the advantage of utilizing the between-family information within the same pedigree, should the identity-by-descent (IBD) coefficients be computed, which could be a very time-consuming task for large family sizes.

Given the information above, within-family analysis may still be useful for dairy cattle QTL analysis purpose because the sample size can be very large (as many as hundreds of offspring for a sire) due to the artificial insemination technology. On the other hand, with data from multiple families available, separate within-family analysis could very well overcome the uncertainty of QTL segregation in founder sires. Since our aims were to better understand the QTL affecting daughter pregnancy rate and other quantitative traits in a previously identified prominent Holstein family, within-family analysis using newly added genetic markers and improved methodology would be meaningful. It was also of interest to check the QTL segregation status of the reproductive traits and milk production traits in other families within the extended pedigree. In this regard, we decided to perform within-family QTL analysis for those large families in our experimental data (see Chapter ?? for our endeavors for across-family analysis).

For a within-family analysis, a sire has at most two variant alleles Q_1 and Q_2 at any loci. It is assumed that the sire is segregating at one or some QTL that determine the trait of interest. Because of the unique feature of (grand) daughter design in which only sire and offspring are genotyped while dams are not, the maternally contributed QTL alleles would not be possible to be traced. However, such maternal effects is well suited to be lumped into the residual because the dams are assumed to be random samples from the dam population. Therefore, the issue at stake is to keep track of the inheritance of sire chromosome (or chromosomal segment) in every offspring. The phenotypic contrast between offspring that inherits a Q_1 allele and those with Q_2 allele would shed light to whether the locus under examination influence the trait and

how large the effect is, which is the basis for QTL mapping in which QTL effects are treated as fixed effects, regardless of how complicated the mathematical formulations could be.

To be able to trace parental alleles or haplotypes, the sire haplotypes need to be inferred first, based on the observed unordered genotypes in the sire and his offspring. In this chapter, we regard this as a solved question and the calculation of conditional probability of QTL is done by using the closest flanking markers whose paternal alleles are definite (see a schematic diagram in Figure 3.1).

In this chapter, we concentrate our efforts on applying the multiple interval mapping model to QTL mapping in a large half-sib family such as that of dairy cattle, in which QTL effects are treated as fixed effects. We also describe in detail of our extensions to the MIM model such as the modeling of heterogeneous residuals, as well as updates to our in-house software Windows QTL Cartographer (WinQTLCart). By applying the new method to a dairy cattle data set from a grand daughter design [MUNCIE *et al.*, 2006], and comparing with the analysis results from the least squares regression method by using QTL Express [SEATON *et al.*, 2002], we demonstrate that our method is more powerful in detecting more suggestive QTL without incurring an elevated significance threshold. Furthermore, besides the comparable results between WinQTLCart and QTL Express such as the suggestive QTL location and the estimate of QTL effect, the 1.5 LOD-drop support interval that can be easily obtained from WinQTLCart outputs, which closely approximates 95% confidence interval [VISSCHER *et al.*, 1996; LAURIE *et al.*, 2008], often gives a much narrower region than the 95% confidence interval obtained through bootstrap re-sampling by QTL Express.

3.2 Single Half-sib Family QTL Analysis: Least Squares Method

Prior to our implementation of (multiple) interval mapping for within-family QTL analysis in a half-sib family, a least squares method proposed by HALEY and KNOTT [1992], also called Haley-Knott regression, is widely in use and the web-interface soft-

ware, QTL Express (<http://qtl.cap.ed.ac.uk/>), is freely accessible to the public. The least squares model for t QTL can be written as below:

$$y_j = \mu + \sum_{r=1}^t \alpha_r \Pr(Q_r | \mathbf{M}_{obs}) + e_j, \quad j = 1, 2, \dots, n \quad (3.2.1)$$

where

- y_j is the trait phenotype of the j -th individual;
- μ is the mean of the trait;
- α_r is the substitution effect of QTL r ;
- $\Pr(Q_{r1} | \mathbf{M}_{obs})$ is the conditional probability of the j -th offspring inherits the Q_{r1} allele from the sire given the observed marker genotypes and marker map information;
- the residual e_j is assumed to be *i.i.d.* distributed as $\mathcal{N}(0, \sigma_e^2)$.

For single QTL detection, the hypothesis testing between zero versus one QTL at a locus can be constructed as either a F-test statistic obtained by taking the ratio between the mean squares of the regression model and the residual, or a log likelihood ratio test statistic between the models under the null and alternative hypotheses. However, the more interesting hypothesis sought for mapping a QTL are H_0 : no QTL on the chromosome versus H_1 : one QTL somewhere on the chromosome. The above test statistic needs to be obtained at each scanned locus and the maximum will be the desired test statistic for the chromosomal scanning for a QTL. The distribution of such test statistic (chromosomal maximum of log likelihood ratios) under the null hypothesis (no QTL on the chromosome) is preferably to be found using data permutations [CHURCHILL and DOERGE, 1994; DOERGE and CHURCHILL, 1996] and an appropriate significance threshold can be obtained from the permutations.

QTL Express, a web-interfaced tool, implements the above regression model and uses F-statistic for the hypothesis testing. It allows users to obtain point-wise, chromosome-wide or experimental-wise significance thresholds from permutations as

well as a 95% confidence interval of the QTL position via bootstrap. QTL Express is also capable of taking into account of varying reliability for the trait by a weighted least squares analysis [SEATON *et al.*, 2002]. However, it becomes awkward and tediously laborious when users want to obtain both a permutation-based significance threshold and a bootstrapped confidence interval with different reliability for different traits to be accounted. Accidental mis-specifications of analysis options during repetitive back-and-forth between QTL Express web pages could cost time; but more potentially dangerously is that such human errors could be difficult to detect and will be inadvertently carried over towards subsequent decision-making.

3.3 Our Method: Weighted Multiple Interval Mapping (wMIM)

In this section, we describe in details our extension of the multiple interval mapping framework [KAO *et al.*, 1999] towards the half-sib data. For simplicity at this time, we consider only the additive QTL effects. The model for t QTL can be specified for a single half-sib family of size n as follows:

$$y_j = \mu + \sum_{r=1}^t b_r^* x_{jr}^* + e_j, \quad j = 1, 2, \dots, n \quad (3.3.1)$$

where

- y_j is the phenotypic value of j -th offspring; μ is the mean of the model;
- b_r^* is the additive effect of the r -th putative QTL;
- x_{jr}^* is an indicator variable for the alleles of the r -th QTL being inherited by the j -th offspring and is defined as follows:

$$x_{jr}^* = \begin{cases} +1/2 & \text{if the QTL allele } Q_{r1} \text{ is inherited;} \\ -1/2 & \text{if the QTL allele } Q_{r2} \text{ is inherited.} \end{cases} ; \quad (3.3.2)$$

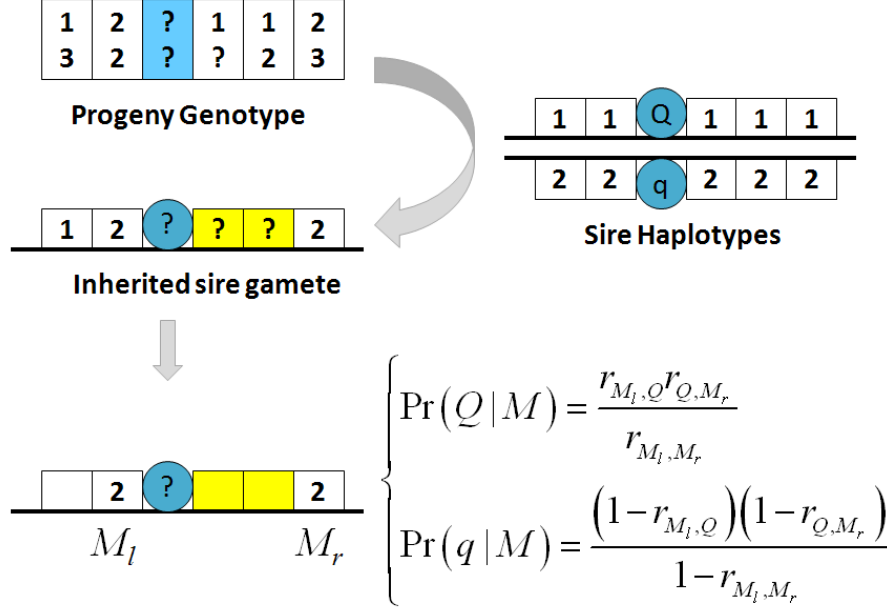


Figure 3.1: Diagram for calculating conditional probability. Only the nearest fully informative markers are used.

- $e_j \sim \mathcal{N}(0, \sigma_e^2 / REL_j)$ where REL_j is the reliability for j -th individual for the trait.

Note that we assume a heteroscedastic residual variance model by incorporating the reliability into the model. If the estimated reliability is not available for some individuals and/or some traits, then REL will be assumed 1.

Statistically the model (3.3.1) is a mixture model because x_{jr}^* , which is the unobserved r -th QTL genotype in j -th offspring, can take different values with respective conditional probabilities. Such conditional probabilities are computed based on the closest pair of flanking markers whose paternal alleles are definite and Figure 3.1 shows a schematic diagram of such calculation.

3.3.1 Maximum likelihood analysis

Maximum likelihood (ML) procedures are used for testing QTL and estimation of QTL effects. The MIM framework is capable of modeling the non-constant residual

variances as specified in equation (3.3.1). First, the log likelihood given the model (3.3.1) can be written as:

$$\log L(\mathbf{E}, \mu, \sigma_e^2 | \mathbf{y}, \mathbf{M}_{obs}) = \sum_{j=1}^n \log \left[\sum_{l=1}^{2^t} p_{jl} \phi(y_j; \mu + \mathbf{D}_l \mathbf{E}, \sigma_e^2 / REL_j) \right] \quad (3.3.3)$$

where \mathbf{D}_l denotes the l -th row of the genetic design matrix \mathbf{D} (as defined in KAO *et al.* [1999]) and the QTL effects vector \mathbf{E} are specified as below:

$$\mathbf{D} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & -\frac{1}{2} \end{bmatrix}_{2^t \times t}, \quad \mathbf{E} = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_t^* \end{bmatrix}_{t \times 1}$$

Secondly, the Expectation-Maximization (EM) algorithm is used for obtaining maximized likelihood and the MLEs, in which the $[s + 1]$ -th E-step gives

$$\pi_{jl}^{[s+1]} = \frac{\pi_{jl}^{[s]} \phi(y_j; \hat{\mu}^{[s]} + \mathbf{D}_l \hat{\mathbf{E}}^{[s]}, (\hat{\sigma}_e^2)^{[s]} / REL_j)}{\sum_{l=1}^{2^t} \pi_{jl}^{[s]} \phi(y_j; \hat{\mu}^{[s]} + \mathbf{D}_l \hat{\mathbf{E}}^{[s]}, (\hat{\sigma}_e^2)^{[s]} / REL_j)} \quad (3.3.4)$$

and the $[s + 1]$ -th M-step involves the following equations:

$$\hat{\mathbf{E}}^{[s+1]} = \text{diag}(\mathbf{V})^{-1} \left\{ (\mathbf{\Pi D})' [\boldsymbol{\gamma} \# (\mathbf{Y} - \hat{\mu}^{[s]} \mathbf{1})] - \text{nondiag}(\mathbf{V}) \hat{\mathbf{E}}^{[s]} \right\} \quad (3.3.5)$$

$$\hat{\mu}^{[s+1]} = \frac{\boldsymbol{\gamma}' (\mathbf{Y} - \mathbf{\Pi D} \hat{\mathbf{E}}^{[s]})}{\boldsymbol{\gamma}' \mathbf{1}} \quad (3.3.6)$$

$$(\hat{\sigma}^2)^{[s+1]} = \frac{[\boldsymbol{\gamma} \# (\mathbf{Y} - \hat{\mu}^{[s]} \mathbf{1})]' (\mathbf{Y} - \hat{\mu}^{[s]} \mathbf{1}) - 2 [\boldsymbol{\gamma} \# (\mathbf{Y} - \hat{\mu}^{[s]} \mathbf{1})]' \mathbf{\Pi D} \hat{\mathbf{E}}^{[s]} + (\hat{\mathbf{E}}^{[s]})' \mathbf{V} \hat{\mathbf{E}}^{[s]}}{(\sqrt{\boldsymbol{\gamma}})' \mathbf{1}} \quad (3.3.7)$$

with

$$\boldsymbol{\gamma} = \begin{bmatrix} REL_1 \\ REL_2 \\ \vdots \\ REL_n \end{bmatrix}_{n \times 1}, \quad \mathbf{V}_{t \times t} = \{\boldsymbol{\gamma}' \boldsymbol{\Pi} (\mathbf{D}^r \# \mathbf{D}^s)\}_{r,s=1,\dots,t},$$

and

$$\boldsymbol{\Pi} = \begin{bmatrix} \pi_{11}^{[s+1]} & \pi_{12}^{[s+1]} & \cdots & \pi_{12^t}^{[s+1]} \\ \pi_{21}^{[s+1]} & \pi_{22}^{[s+1]} & \cdots & \pi_{22^t}^{[s+1]} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{n1}^{[s+1]} & \pi_{n2}^{[s+1]} & \cdots & \pi_{n2^t}^{[s+1]} \end{bmatrix}_{n \times 2^t},$$

where \mathbf{D}^r denotes the r -th column of the matrix \mathbf{D} and $\#$ is the Hadamard product. Initial values to start the E-M iterations are usually chosen to be simple moment estimators and iterations will be continued until convergence criteria are met [WANG *et al.*, 2007].

Note the highlighted vector $\boldsymbol{\gamma}$ of REL of each individual for the trait that has nicely fitted into the E-M equations for MIM. When REL data is not available, $\boldsymbol{\gamma}$ will be a vector of 1s such that this wMIM reduces to MIM [KAO *et al.*, 1999]. Such an extension to allow heteroscedastic residual variances expands the applicability of MIM and WinQTLCart without incurring difficulty in interpretation of the estimate of QTL effect.

3.3.2 Hypothesis testing

Once the ML procedure is completed and maximum likelihood estimates (MLEs) of parameters have been obtained, under a hypothesis such as $H_0 : b_r^* = 0$ versus $H_1 : b_r^* \neq 0$, which tests whether there is r -th QTL given the previous $r - 1$ QTL

having been detected, the likelihood ratio (LR) test is constructed as follows:

$$\begin{aligned}
 LR &= -2 \ln \frac{\operatorname{argmax}_{\Theta \in H_0} L(\mu, b_1^*, \dots, b_t^*, \sigma_e^2 | data)}{\operatorname{argmax}_{\Theta \in H_0 \cup H_1} L(\mu, b_1^*, \dots, b_t^*, \sigma_e^2 | data)} \\
 &= -2 \ln \frac{L(\tilde{\mu}, \tilde{b}_1^*, \dots, \tilde{b}_{r-1}^*, \tilde{b}_r^* = 0, \tilde{b}_{r+1}^*, \dots, \tilde{b}_t^*, \tilde{\sigma}_e^2)}{L(\hat{\mu}, \hat{b}_1^*, \dots, \hat{b}_t^*, \hat{\sigma}_e^2)}
 \end{aligned}$$

where $\hat{\mu}, \hat{b}_1^*, \dots, \hat{b}_t^*, \hat{\sigma}_e^2$ are MLEs of $\mu, b_1^*, \dots, b_t^*, \sigma_e^2$ under the alternative hypothesis H_1 and $\tilde{\mu}, \tilde{b}_1^*, \dots, \tilde{b}_{r-1}^*, \tilde{b}_{r+1}^*, \dots, \tilde{b}_t^*, \tilde{\sigma}_e^2$ are MLEs of $\mu, b_1^*, \dots, b_{r-1}^*, b_{r+1}^*, \dots, b_t^*, \sigma_e^2$ under the null hypothesis H_0 with b_r^* constrained to zero. LOD (logarithm of odds) score can be easily obtained by the equation $\text{LOD} = 0.217 \text{ LR}$.

The chromosome-wide significance threshold values can be obtained via permutations CHURCHILL and DOERGE [1994]; DOERGE and CHURCHILL [1996] for detecting the first QTL. According to LAURIE *et al.* [2008] and SILVA [2009], we can either use this threshold for next-step testing or a score re-sampling procedure to obtain thresholds at each step of adding a QTL or not [ZOU *et al.*, 2004]. We will use the permutation LR threshold in this chapter because most traits in the dairy cattle data have only one QTL in the interested chromosomal region.

3.3.3 Calculation of R^2

Parameter estimates are obtained at the suggestive QTL and the predicted trait values are computed as a weighted sum:

$$\hat{y}_j = \hat{\mu} + \sum_{k=1}^2 \hat{\pi}_{jk} x_{jk} \hat{b} \quad (3.3.8)$$

where $\hat{\mu}$ and \hat{b} are estimates of the mean trait value and the QTL additive effects, x_{jk} is the indicator value of the k -th genotypes of the QTL in the individual j while $\hat{\pi}_{jk}$ is the estimates of the posterior probability. Then the coefficient of determination,

R^2 , is computed by

$$R^2 = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2} \quad (3.3.9)$$

which tells the portion of total variations that is due to QTL.

3.3.4 Windows QTL Cartographer software updates

Our in-house software – Windows QTL Cartographer [WANG *et al.*, 2007] has been updated, with the help from Dr. Shengchu Wang (<http://statgen.ncsu.edu/~shchwang/>), to accommodate this extensions as well as the file format revisions. Now it is capable of analyzing data from HS-SF (half-sib single family) designs with or without specifications for heteroscedastic residual variances. Interval mapping (IM) and multiple interval mapping (wMIM) can be performed for each trait and each family. Test thresholds under an appropriate chromosome-wide significance level can be obtained from permutations for each trait, respectively. Note that in case of heteroscedastic residual models, the permutation procedure was implemented in such a way that the association between traits and the respective reliability is preserved over permutations; instead, the association between genotypes and phenotypes and reliability, if there is, is broken and permuted.

3.4 Application of wMIM to Experimental Data

Extensive simulation studies have already been performed to assess the MIM methodology [KAO *et al.*, 1999; LAURIE *et al.*, 2008]. In particular, KAO [2000] conducted a comprehensive comparison between the least squares regression method and the mixture model-based MIM approach and showed that the MIM framework is more powerful and gives better parameter estimates especially the residual variances. Therefore, we opted to directly apply the wMIM method towards a dairy cattle half-sib data and compare the results with that of the Haley-Knott regression method.

The experimental data to which we apply our newly extended method are from Dr. Melissa Ashwell's lab (Department of Animal Science of North Carolina State University). Previous QTL analysis has been reported by ASHWELL *et al.* [2004] and MUNCIE *et al.* [2006] in which suggestive QTL for traits such as daughter pregnancy rate (DPR), productive life (PL) and somatic cell score (SCS) had been found in a 60cM segment on BTA 18. Additional microsatellite markers were then genotyped in a promising family as well as families descending from that family through paternity. Animal Improvement Program Laboratory (AIPL) records of these animals for more traits such as milk yield (MY), fat yield (FY) and percentage (FP), protein yield (PY) and percentage (PP), calving ease (CE) and percent difficult births (PDB) were also obtained, alongside their respective estimated reliability information.

3.4.1 Data explorative analysis

Explorative analysis of dairy cattle data before QTL mapping analysis is to examine, detect and correct errors and missing values, and more importantly, to assess the features of the data and justify choice of modeling specifications.

Resource populations

Approximately 930 male descendants from the same Holstein sire were selected as previously described in MUNCIE *et al.* [2006]. This complex pedigree consists of 3 generations with 13 half-sib families, all descending from the same sire [MUNCIE *et al.*, 2006]. For single family analysis, we use eight families each with more than 30 sons (Table 3.4.1) and the total number of animals was 834. Their family sizes range from 46 to 176, with an average number of 104 sons.

Marker map and marker genotypes summary

A total of 23 microsatellite markers spanning a 59.6cM segment (Table 3.4.1) on BTA 18 were genotyped. The founder sires of different families have different subsets of segregating markers and the average intervals between segregating markers vary

Table 3.1: Marker map summary of the chromosomal segment on BTA 18.

Family	Number of Offspring	Number of Segregating Markers (cM)	First Segregating Marker (cM)	Last Segregating Marker (cM)	Segment Length (cM)	Average Interval (cM)
I-1	168	17	24.5	84.1	59.6	3.7
II-1	66	16	31.5	84.1	52.5	3.5
II-2	176	11	24.5	76.8	52.3	5.2
II-5	63	18	24.5	84.1	59.6	3.5
II-6	109	15	24.5	84.1	59.6	4.3
III-1	80	13	24.5	84.1	59.6	5.0
III-3	126	12	24.5	84.1	59.6	5.4
III-6	46	13	24.5	84.1	59.6	5.0

among founder sires in the range of 3.5cM and 5.4cM, with an overall average interval of 4.3cM and a standard deviation of 4.2cM across all the 8 founder sires (Table 3.4.1). The large standard deviation of the interval lengths is caused by the uneven spacing between markers (Table 3.4.1) and the varying set of segregating markers in sires (Table 3.4.1). All markers are polymorphic except BMS929 (Table 3.4.1) and failures in genotyping were under control for most of markers except the marker DIK5222. However, because of the high density of genotyped markers in its vicinity and more importantly, because our method of sire haplotype reconstruction is robust against missing markers and the calculation of conditional probabilities makes use of the nearest informative marker available in an individual, DIK5222 was not excluded.

Most of markers show an observed heterozygosity in the range of 0.6 – 0.8 (Table 3.4.1). There were five markers (shown with asterisks in Table 3.4.1) that were selectively genotyped in only two families: I-1 and III-3. Their respective missing value percentage and the observed heterozygosity were computed from those two families only. Not all markers are segregating in all founder sires, which is shown in both Tables 3.4.1 and 3.4.1. However, the marker coverage areas as well as the average inter-marker intervals do not vary much across founder sires (Table 3.4.1). In our QTL analysis, only those segregating markers in founder sires were used for inference of sire haplotypes and computation of conditional probabilities.

Table 3.2: Microsatellite marker data summary. The purple asterisks (*) represent markers that were genotyped in families I-1 and III-3 only and their percentages of missing data and observed heterozygosities were computed within these two families only.

Markers	Position (cM)	Missing (%)	Number of Alleles	Heterozygosity
BMS2213	24.487	12.0	8	0.8300
INRA121	30.153	3.1	8	0.6752
DIK1112*	30.685	8.2*	3	0.3651*
DIK2848*	31.543	13.4*	3	0.6303*
DIK4792	32.076	18.8	5	0.8129
BR4406	33.397	20.5	5	0.2945
BM8151	40.213	4.6	6	0.3599
DIK2128	40.213	7.5	7	0.7214
DIK5222	41.296	55.0	4	0.6702
DIK4896*	42.136	20.6*	6	0.6286*
HAUT14	42.136	14.4	7	0.7545
DIK4059*	44.009	27.7*	4	0.5506*
MNB27	44.009	11.0	12	0.7076
BM7109	46.976	1.0	6	0.6643
INRA063	47.953	16.9	4	0.5443
BMS2914	50.041	5.2	6	0.6291
DIK2738*	51.952	9.4*	7	0.2663*
ILSTS002	54.173	7.2	5	0.7657
DIK2779	56.311	10.6	5	0.4781
BMS929	61.157	1.0	2	0.5923
BB710	62.089	3.4	5	0.7269
BM2078	76.782	8.1	8	0.7933
TGLA227	84.087	11.8	10	0.8708

Information contents for each family are shown in Figure 3.2, calculated from variances of QTL conditional probabilities at each centiMorgan as a proportion of the variance when true descent is known [SPELMAN *et al.*, 1996]. At most positions, the profiles are generally flat and high in the range of 0.8 – 1.0. The low level of information content between 40 and 45cM for families III-1 and III-6 results from the following facts: (i) there are only a few informative markers for these two families in this region, namely HAUT14, MNB27, and ILSTST002 (shown in Figure 3.2); (ii) the paternal alleles transmitted from the sire to offspring at these three markers happen to be almost unequivocally identical among offspring within the same family. Therefore, the QTL conditional probabilities are highly homogeneous among these offspring, which leads to low variances of conditional probabilities thus the low information contents observed for these two families.

Phenotypes and reliability

Phenotypes were collected from at least 50, and in many cases, thousands of daughters for each bull and PTA or DD/DYD trait values were obtained via the animal model genetic evaluation [WIGGANS and VANRADEN, 1989; VANRADEN and WIGGANS, 1991]. Figure 3.3 shows that phenotypic data for most of traits has different mean values but similar scales of variations across families. Histograms of all the trait data for each family were also checked (not shown) and only a few possible outliers were noticed but not excluded from our analysis.

Table 3.4.1 summarizes the phenotypic data across families in which we found that the means and medians match well with each other; what's more, the skewness coefficient and kurtosis are mostly around zero, suggesting that skewness and kurtosis are not significant problems with most of these trait data, as was also observed from those histograms. Since the distribution of trait values within genotypic marker classes follows a mixture distribution, and because of the expectation that there are QTL, the skewed distribution for some traits is not surprising and we worked with untransformed data. Very few individuals have missing values in phenotypic data and

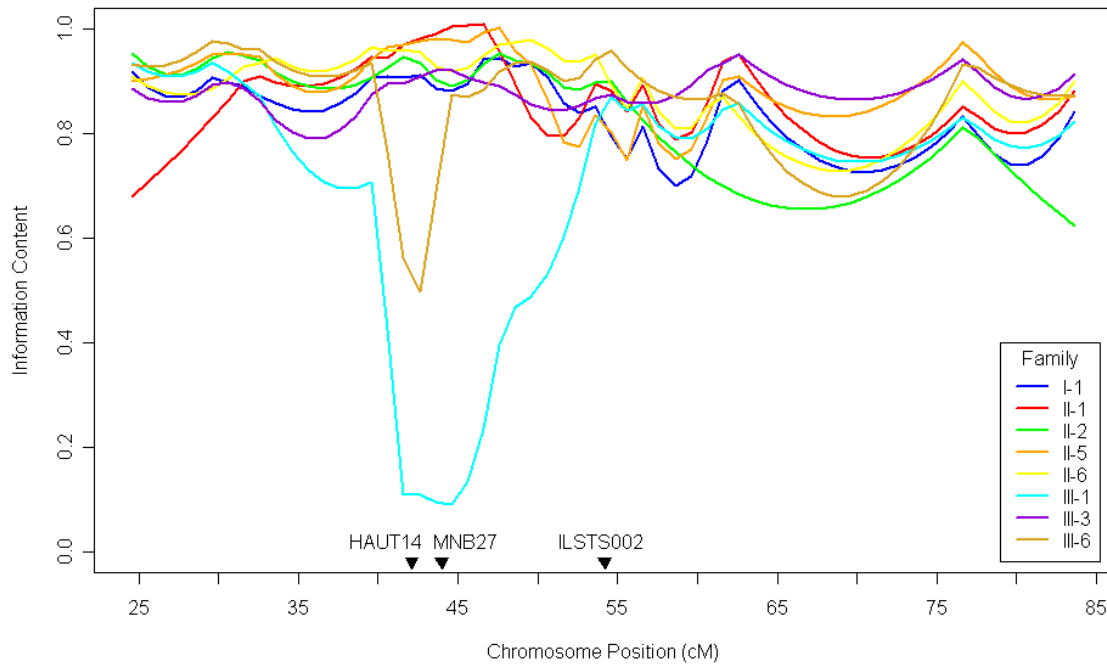


Figure 3.2: Information content derived for the marker coverage region for each family. Information content was calculated from variances of QTL conditional probabilities at each centiMorgan as a proportion of the variance when true descent. All profiles are in the range of 0.8 – 1.0 except the ones between 40 and 45cM for families III-1 (cyan) and III-6 (golden). Positions of the informative markers contributing to the observed low information content in families III-1 and III-6 are also shown.

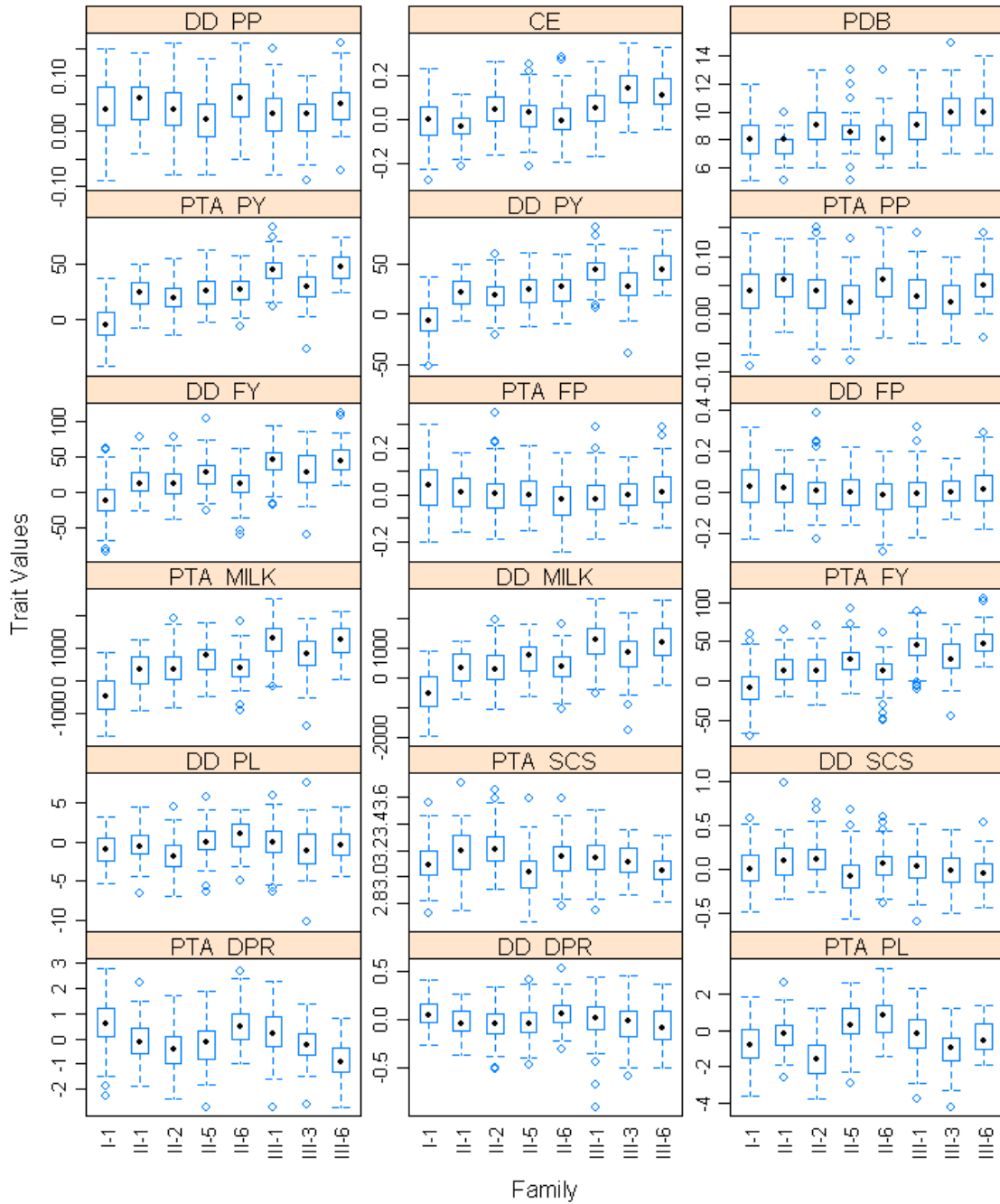


Figure 3.3: Box plot of each trait within each family. The black dots represent the median values and the rectangles represent the inter-quartile regions (IQR, between quartiles of 25% and 75%).

Table 3.3: Phenotypic data summary. The skewness coefficient and kurtosis are mostly around zero, suggesting that neither is a significant problem with most of these trait data.

No.	Trait (unit)	Missing (%)	Median	Mean	Std Dev	Min	Max	Skew -ness	Kurt -osis
1	PTA DPR	0.2	0	0	0.1	-0.3	0.4	0.3	0.5
2	DD DPR	0.2	19	18.1	29.8	-83	112	-0.1	0.1
3	PTA PL	0	0	0	0.1	-0.3	0.3	0.2	0
4	DD PL	0.1	0	0	0.2	-0.9	0.5	-0.3	1
5	PTA SCS	0.2	439	446.6	829.7	-1966	2698	-0.1	-0.2
6	DD SCS	2.9	-0.7	-0.7	2.1	-10.1	7.4	-0.1	0.3
7	PDB (%)	0.2	0	0	0	-0.1	0.2	-0.1	-0.1
8	CE	0.2	23	21.6	22.9	-52	87	-0.2	0
9	PTA MILK (lb)	1.5	0	0	0.2	-0.6	1	0.2	0.4
10	DD MILK (lb)	0	9	8.7	1.6	5	15	0.5	0.4
11	PTA PY (lb)	0	0	0	0.9	-2.7	2.8	0	-0.1
12	DD PY (lb)	0	0	0	0.1	-0.2	0.4	0.4	0.5
13	PTA PP (%)	0	20	18.8	27.3	-71	106	-0.1	0.1
14	DD PP (%)	0	485	472.7	772.4	-1723	2563	-0.2	-0.3
15	PTA FY (lb)	0	-0.5	-0.6	1.3	-4.2	3.4	0	-0.2
16	DD FY (lb)	0	0	0	0	-0.1	0.2	-0.1	0
17	PTA FP (%)	0	25	22.8	21.6	-43	84	-0.3	-0.1
18	DD FP (%)	0	3.1	3.1	0.2	2.7	3.7	0.1	0.4

those missing values were ignored in our analysis.

In examining the reliability data, we found that the reliability values for different traits and for different offspring within the same family vary greatly (Figure 3.4). This called for the need to incorporate reliability into modeling the residual variances in our MIM framework (Equation 3.3.1) and we call it the weighted MIM (wMIM). We compared the results between analyses using and ignoring reliability information and found that wMIM approach is more powerful in that it gives higher LOD values without incurring a higher chromosome-wide significance threshold (results not shown). All suggestive QTL found by MIM analysis without using reliability data were also detected in the wMIM analysis under the same significance level, but not vice versa. Therefore, we only present results from wMIM analysis in this report.

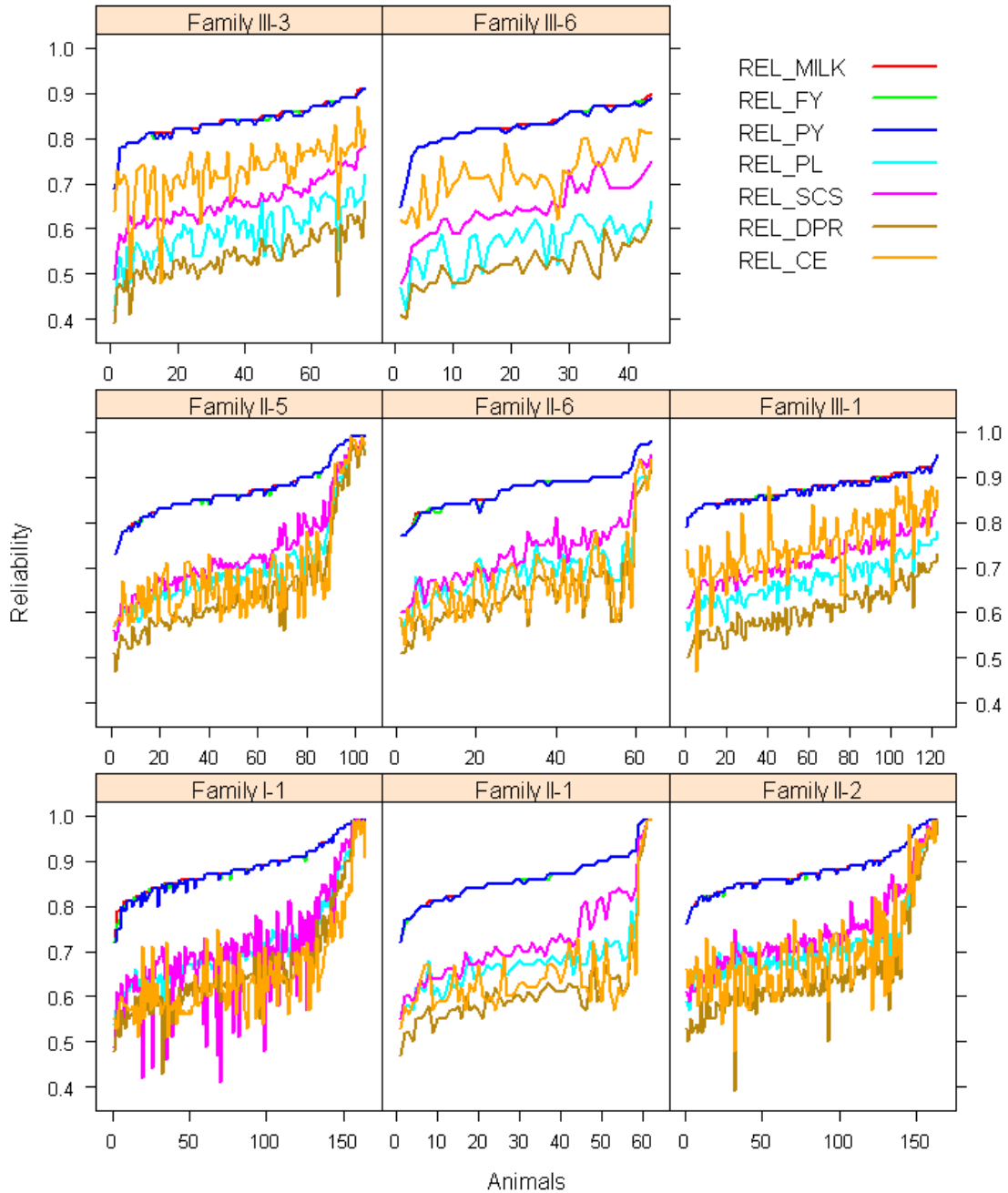


Figure 3.4: Reliability of different traits. Only families with more than 30 offspring are shown. Animals are sorted by the reliability for milk yield.

3.4.2 Within-family QTL analysis results

We first performed interval mapping for each trait within each family, separately, using Windows QTL Cartographer (WinQTLCart) [WANG *et al.*, 2007]. The threshold for LOD score, under chromosome-wide significance levels of 0.05 and 0.01, were obtained via 2500 permutations. There are twenty suggestive ($p < 0.05$) QTL for eight traits detected in six families, respectively, summarized in Table 3.4.2. Among them, nine are significant ($p < 0.01$) QTL that exceed the threshold under the chromosome-wide significance level of 0.01. The estimated QTL effects are in their absolute values because their signs are arbitrary depending on the uncertain assignment of QTL alleles in a family. The LOD profiles of those suggestive QTL are plotted in Figure 3.5.

For both verification and evaluation purposes, we compared our results from WinQTLCart with those from QTL Express and found that for most traits, both tools give comparable results such as whether and where there would be a suggestive QTL, the estimate of QTL substitution effect, etc. However, the 1.5 LOD-drop support interval often gives a much narrower region than the 95% confidence interval obtained through bootstrap re-sampling by QTL Express (results not shown). In addition, WinQTLCart was able to identify a few more suggestive QTL that QTL Express did not under the same level of significance level (results not shown). Therefore, we will only refer to our WinQTLCart results below.

There is a suggestive QTL for DPR in family I-1 in the range of 42 and 55cM, as detected by wMIM analysis on both PTA and DD DPR trait data. In family II-1, a suggestive QTL was also detected with a peak at 55cM but with a much wider support interval ranging from 32 to 84cM.

There is a suggestive QTL detected for PTA PL in family III-3 at approximately 35cM with the 1.5 LOD support interval ranging from 25 to 42cM. The same QTL was also detected for DD PL at the same peak location but with a much wider support interval in the same family. Suggestive QTL for PL trait were also detected in families I-1 and III-6 with peaks at 50 and 55cM, respectively. The different patterns of LOD

Table 3.4: Summary of suggestive QTL from within-family QTL interval mapping analysis. [‡], significant at the chromosome-wide level of 0.01.

Trait	Family (Size)	Max LOD	95% LOD Threshold	Peak (cM)	1.5-LOD Support Interval (cM)	Absolute QTL Effect	R^2 (%)
PTA PL	I-1 (168)	1.78	1.72	49.6	24.6 - 64.6	0.49	5.2
PTA PL [‡]	III-3 (126)	4.41	1.43	34.6	24.6 - 41.6	0.94	18.3
DD PL [‡]	III-3 (126)	3.08	1.47	35.6	24.6 - 82.6	1.43	12.2
DD PL	III-6 (46)	1.99	1.59	54.6	24.6 - 65.6	1.61	18.2
PTA DPR [‡]	I-1 (168)	4.43	1.74	49.6	42.6 - 54.6	0.60	12.8
DD DPR [‡]	I-1 (168)	3.79	1.60	44.6	41.6 - 53.6	0.10	10.8
DD DPR	II-1 (66)	1.49	1.49	54.6	31.6 - 83.6	0.09	10.0
DD SCS	I-1 (168)	1.78	1.65	24.6	24.6 - 53.6	0.09	5.3
DD SCS	II-5 (63)	2.18	1.64	39.6	24.6 - 53.6	0.18	15.3
PTA SCS [‡]	II-5 (63)	2.73	1.76	38.6	24.6 - 52.6	0.15	19.8
PDB	II-2 (176)	1.53	1.43	54.6	24.6 - 76.6	0.56	4.3
PDB [‡]	III-6 (46)	2.66	1.70	75.6	62.6 - 83.6	1.62	25.7
CE [‡]	III-6 (46)	2.49	1.64	76.6	31.6 - 83.6	0.09	22.6
PTA MILK [‡]	II-2 (176)	2.36	1.39	24.6	24.6 - 42.6	268.38	6.7
DD MILK	II-2 (176)	2.30	1.49	24.6	24.6 - 40.6	286.02	6.4
PTA PY [‡]	II-2 (176)	2.42	1.45	24.6	24.6 - 44.6	6.65	6.6
DD PY	II-2 (176)	2.03	1.37	24.6	24.6 - 76.6	6.65	5.5
DD FY	I-1 (168)	1.76	1.59	53.6	24.6 - 83.6	12.29	5.7
PTA FY	I-1 (168)	1.82	1.60	53.6	24.6 - 83.6	11.24	5.9
PTA FY	II-2 (176)	1.38	1.36	24.6	24.6 - 76.6	7.14	3.7

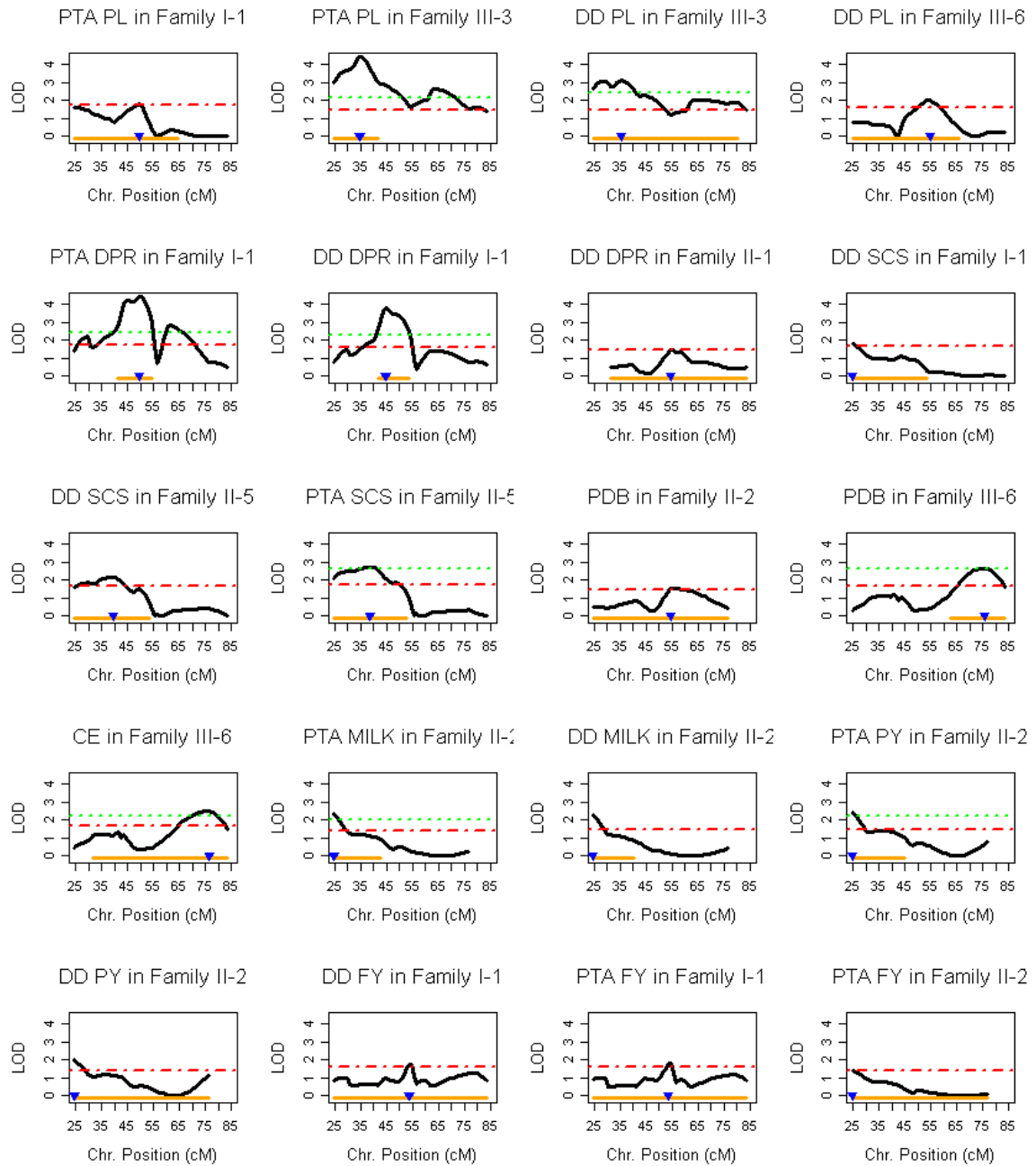


Figure 3.5: QTL interval mapping results. Red dashed lines and green dotted lines are thresholds under the chromosome-wide significance level of 0.05 and 0.01, respectively. Triangles (blue) point to locations of the maximum LOD values and the horizontal segments (orange) at the bottom of graphs represent the 1.5 LOD support interval for QTL.

profiles of PL trait among different families (I-1, III-3 and III-6), however, may suggest different QTL segregating in these families.

A suggestive QTL for SCS were located between 25 and 54cM with the peak at 39 – 40cM in family II-5, as detected for both PTA and DD SCS trait data. The LOD profiles for PDB and CE in family III-6 were similar in pattern with almost identical location of the peaks. The 1.5 LOD support interval for PDB in family III-6 was from 63 to 84cM, much narrower than that of CE in the same family (32 – 84cM). In family II-2, there was also a suggestive QTL for PDB with LOD peak located at 55cM, though the signal was barely strong enough for the peak to exceed the 95% chromosome-wide significance threshold.

Suggestive QTL for yield traits were also detected in this 60cM segment on BTA 18. The suggestive QTL for milk yield in family II-2 was located on the boundary of marker coverage region: 25cM. The 1.5 LOD-drop support interval for this suggestive QTL extends to the position of 41cM for DD Milk trait data. Note that the LOD profiles between PTA and DD values were almost identical, which were also observed for other yield traits such as protein yield (PY) in family II-2 and fat yield (FY) in family I-1, respectively. In addition, there was a suggestive QTL for FY detected in family II-2.

For a few traits such as PTA PL and DD PL in family III-3 and PTA DPR in family I-1, there were multiple LOD peaks that all exceeded the 95% chromosome-wide significance threshold from the weighted IM analysis using WinQTLCart (Figure 3.5). To determine whether there were indeed multiple QTL involved, the multiple interval mapping (MIM) analysis was performed using WinQTLCart. The MIM analysis did not find a significant second QTL when the first QTL was included in the model (results not shown). Therefore, we concluded that Table 3.4.2 summarizes all the detected QTL from the within-family ML-based QTL analysis.

3.4.3 Discussions

We updated WinQTLCart so that it now can analyze the outbred single half-sib family data. Our analysis has thus supported the existence of QTL for traits such as

the daughter pregnancy rate and productive life in the interested region on bovine chromosome 18. It also detected significant ($p < 0.01$) or suggestive ($p < 0.05$) QTL for a number of other traits such as somatic cell score, calving ease, percent difficult births and the milk composition traits (milk, fat and protein yields) in the same chromosomal region on BTA 18.

Multiple interval mapping analysis is capable of modeling multiple putative QTL in the same model simultaneously and has been shown to be more powerful [KAO *et al.*, 1999] if there indeed exists more than one QTL. We checked every trait where there seemed to be more than one peak exceeding the 95% chromosome-wide significance threshold from the interval mapping analysis using the updated MIM module in WinQTLCart and found only single suggestive QTL in all the cases. However, the capability of WinQTLCart for performing MIM analysis on single half-sib family data is noteworthy and will be useful when there is indeed more than one QTL.

In comparison with QTL Express, there are following advantages that WinQTLCart offers for within-family QTL analysis for half-sib populations:

1. WinQTLCart implements the maximum likelihood method that has been shown to outperform the regression interval mapping implemented by QTL Express in terms of both the testing power and the accuracy and unbiasedness of the estimates [KAO, 2000];
2. WinQTLCart offers a user-friendly graphical interface and allows batch analysis of multiple traits to be performed, especially when each trait has its own reliability values;
3. WinQTLCart produces comprehensive outputs of analysis results such as the LOD profiles over the genomic region, the estimates of the QTL effect and R^2 at every putative position, the significance threshold values, etc.;
4. WinQTLCart offers well-formatted outputs so that users can opt to retrieve the desired analysis results such as the values and the locations of LOD peaks and the LOD-drop support intervals by parsing the output files in a convenient

manner, which minimizes the chance of accidental human errors during post-analysis parsing and summarization of results.

3.5 Prospectives

3.5.1 Further upgrades to Windows QTL Cartographer

Although WinQTLCart now has a new function module to analyze HS-SF (half-sib single family) data, the user interface can be further improved such as a dedicated drop-down menu for outbred analysis, and some useful functions so as to detect pedigree and genotype errors can be integrated. Errors in pedigree could arise from inaccurate record of paternity and maternity or accidental human errors such as mis-specification of genders or conflicting records of the same individual, etc. Genotype errors refer to conflicting marker genotypes among related individuals according to Mendelian rules, which could be caused by contaminated DNA samples, mutations or inaccurate genotyping. It will be useful and convenient if all such conflicting genotypes can be well documented in a log file, allowing users to examine and correct each conflicting genotype conveniently and in a context-dependent manner.

The score re-sampling procedure for obtaining significance threshold during step-wise model selection of multiple QTL models has already been implemented in WinQTLCart for inbred designs [LAURIE *et al.*, 2008]. However, this functionality has not been connected to the HS-SF module, although such adaptation would be quite straightforward.

Another direction is to extend from single trait QTL analysis to joint analysis of multiple traits. It is well known that various traits such as milk yield, fat yield and protein yield are highly correlated and a joint analysis of these traits would offer a better power performance in dissecting the underlying genetic architectures. Multiple traits multiple interval mapping (MT-MIM) is being studied by Luciano Silva [SILVA, 2009] and the important issue of testing pleiotropy versus close linkage would be of great potential for better understanding major genes influencing one or a few related quantitative traits of economical importance. In the future, MT-MIM method can

also be extended to outbred single families without much modifications.

3.5.2 Multiple family QTL analysis as fixed effects

In a similar way as the Haley-Knott least squares method being extended to multiple unrelated half-sib families [HOESCHELE *et al.*, 1997; HOESCHELE, 2007], our wMIM can also be extended to analyze multiple families with additive QTL effects considered as fixed effects:

$$y_{ij} = \mu_i + \sum_{r=1}^t a_{ri} x_{ijr}^* + e_{ij} \quad (3.5.1)$$

where m is the number of half-sib families and n_i is the number of individuals in i -th family; t is number of QTL in the model and a_{ri} is the QTL additive effects of the r -th QTL; x_{ijr}^* denotes the indicator genotype at the r -th locus of j -th offspring in family i ; e_{ij} is the residual assumed to be independently distributed as a normal distribution with mean zero but a family- and individual-specific variances σ_i^2/REL_{ij} , if REL is available.

The maximum likelihood procedures using an E-M algorithm for the model above can be readily derived (see Appendix A).

Acknowledgement

The author thanks Dr. Shengchu Wang for his great assistance in making software upgrades to Windows QTL Cartographer as well as revisions to the .MCD file format definitions.

Chapter 4

Across-family Mixed Model Analysis of QTL

4.1 Introduction

As already discussed in the introduction of Chapter 3, it is more reasonable to assume more than two alleles segregating at the QTL across multiple families, especially for outbred populations. Also due to the divergent genetic background between outbreeding individuals, the paternal and maternal alleles at a locus may have varying contrasts with respect to the trait phenotypes under their influence. Therefore, it would be more meaningful to assess the variation caused by QTL instead of the allele contrast(s).

The least squares method (LS) [HALEY and KNOTT, 1992; MARTINEZ and CURNOW, 1992] can be used for multiple family QTL analysis for a variety of populations: crosses between multiple inbred lines, outbred half-sibs and full-sibs, etc. In its model, QTL effects are treated as fixed effect such that one QTL effect parameter is usually needed per allele substitution. When it is of interest to test whether a locus affects the trait of interest, all the QTL effect parameters at that locus, one per family, need to be jointly tested. The numerator degree of freedom (usually the number of parameters being tested) of the test statistic will thus be large if the number of families (or crosses between inbred lines) is large, which would be detrimental to the statistical testing

power [SNEDECOR and COCHRAN, 1989].

On the other hand, variance components (VC) models would be appropriate to handle data consisted of multiple or many inbred crosses or outbred families, because they make no specific assumptions about the number of QTL alleles. VC approach is more appropriate when there are a large number of small families like that of human populations. In VC models, QTL effect is treated as a random effect and often assumed to be normally distributed with mean zero and a scalar variance parameter (say σ_Q^2). The correlation structure for the QTL is usually modeled by the identity-by-descent (IBD) matrix which represent the proportion of alleles at a locus that two individuals actually share IBD. In addition, VC models contain two other variance components: the polygenic effects whose correlation structure is captured by the kinship coefficients and the random residual. The question of QTL mapping now becomes to test whether the variance component due to a putative QTL would be non-zero. The readily available estimates of variance components for QTL, polygenic and residuals also make it handy to compute heritabilities. QTL mapping using VC approach has been developed independently by both human genetics [AMOS, 1994; ALMASY and BLANGERO, 1998] and animal breeding [GRIGNOLA *et al.*, 1996; GEORGE *et al.*, 2000] communities.

In dairy cattle breeding, a bull that shows strong trait performance is usually assessed in preliminary analysis as of whether there is genetic control concerning the trait performance. Therefore, when it comes to QTL mapping using both genetic genotypes information from the bull (sire) and his much expanded family of half-sib offspring (sons if it is the grand daughter design or daughters if daughter design), it would not be a concern whether the QTL being sought is segregating in that sire. However, for several or many half-sib families, each with strong segregation in the trait phenotype, there is no way to ascertain the segregation status of the same locus across sires. Therefore, QTL effects would be better suited to be treated as random effects.

In this chapter, we report our efforts of across-family joint QTL analysis by adopting a mixed effects modeling approach, in which QTL effects are assumed to be normally distributed with non-zero mean such that we have both fixed and random

effects terms for a QTL. Because the model specifications include both fixed and random QTL effects, there are several ways to formulate the testing hypothesis - testing only on the fixed or random effects or both. We can reject the null hypothesis of no QTL at a putative locus if either or both the mean and variance parameters of QTL effects are non-zero, because there could be situations where the non-zero mean of QTL effects accounts for the main signal in the data. Our mixed model approach was applied to a large dairy cattle data from a grand daughter design and higher or at least comparable hypothesis testing performance with respect to VC approach was observed. With regard to the alternative testing hypothesis constructions, our mixed model analysis further detected some QTL fixed effects that are not observed in either LS or VC analysis, which may be worthy of further elucidation.

4.2 Variance Components Approach

To facilitate description and discussion of our method for across-family QTL analysis, we'd like to review the framework of variance components models first. A single QTL variance component model can be written as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_q\mathbf{q} + \mathbf{Z}_u\mathbf{u} + \mathbf{e} \quad (4.2.1)$$

where

- \mathbf{y} is the vector of trait phenotype data;
- $\boldsymbol{\beta}$ is a vector of fixed effects, \mathbf{X} is a design matrix relating the appropriate fixed effects to each individual;
- \mathbf{u} is a vector of random polygenic additive effects;
- \mathbf{Z}_u is an incidence matrix relating the appropriate random polygenic effects to each individual;
- \mathbf{q} is a vector of additive QTL effects at a locus of interest;

- \mathbf{Z}_q is an incidence matrix relating individual's trait phenotype with random QTL effects vector;
- \mathbf{e} is a vector of random residuals assumed to be distributed as $\mathcal{N}(0, \sigma_e^2 \mathbf{I})$.

The variance components according to the model (4.2.1) can be derived readily as $Var(\mathbf{y}) = \mathbf{V} = \sigma_q^2 \mathbf{Z}_q \mathbf{G} \mathbf{Z}_q' + \sigma_u^2 \mathbf{Z}_u \mathbf{A} \mathbf{Z}_u' + \sigma_e^2 \mathbf{I}$, where the correlation matrix for polygenic additive effects, \mathbf{A} , equals to twice of the kinship coefficient matrix based on pedigree relationship only and the correlation matrix for QTL additive effects, \mathbf{G} , is the IBD coefficients matrix estimated from observed marker data. If the IBD matrix is estimated at a locus by using all markers on the chromosome, such estimation is called multipoint IBD estimation and QTL mapping using such IBD matrices is called multipoint mapping. Since the incidence matrices \mathbf{Z}_q and \mathbf{Z}_u are usually the identity matrix, the variance components can be further simplified to $\mathbf{V} = \sigma_q^2 \mathbf{G} + \sigma_u^2 \mathbf{A} + \sigma_e^2 \mathbf{I}$. The total genetic and QTL heritabilities are there computed as $h_{total}^2 = (\sigma_q^2 + \sigma_u^2) / (\sigma_q^2 + \sigma_u^2 + \sigma_e^2)$ and $h_{QTL}^2 = \sigma_q^2 / (\sigma_q^2 + \sigma_u^2 + \sigma_e^2)$, respectively. The fraction of genetic variations due to QTL can be computed as $F_{QTL} = \sigma_q^2 / (\sigma_q^2 + \sigma_u^2)$.

The testing hypothesis of whether there is a QTL versus no QTL at the test location (d) can be specified as $H_0 : \sigma_q^2 = 0$ vs. $H_1 : \sigma_q^2 > 0$. A likelihood ratio test (LRT) statistic can be constructed from the two models under the null and alternative hypothesis, respectively. Under the null hypothesis, such a test statistic follows a mixture distribution of half χ_0^2 and another half χ_1^2 [SELF and LIANG, 1987]. However, in the chromosome/genome scanning for a QTL, the distribution of the maximum log likelihood ratio under the null hypothesis that there is no QTL anywhere on the chromosome/genome is unknown, yet GEORGE *et al.* [2000] suggested that it can be approximated by a χ_1^2 distribution. The widely used permutation method [CHURCHILL and DOERGE, 1994; DOERGE and CHURCHILL, 1996] for obtaining the chromosome-wide or genome-wise significance threshold is not applicable for general pedigrees, because it is yet unclear how to accomplish a permutation such that the association between QTL and trait is destroyed but the association between polygenic effect and the trait remains intact. What's more, even for a simple pedigree for which permutations would be valid, the time-consuming IBD calculations for each permuted

data along with the REML analysis of the variance component model would make it practically impossible to obtain a permutation threshold.

4.3 Our Approach: Mixed Effects Model

4.3.1 Mixed effects model specifications

Assuming families are unrelated, we use a mixed effects linear model to decompose the QTL effects into both fixed and random effects, accounting for the different QTL segregating status of QTL across families. A single QTL model can be written as follows:

$$y_{ij} = \beta_0 + \beta_1^* x_{ij}^* + b_{i0} + b_{i1}^* x_{ij}^* + e_{ij} \quad (4.3.1)$$

- y_{ij} is the phenotypic value of j -th offspring of the sire i ;
- β_0 and β_1^* are the overall intercept and QTL fixed effect;
- x_{ij}^* is the QTL genotype indicator and is defined as $x_{ij}^* = +1/2$ if the QTL allele Q_{i1} is inherited, and $-1/2$ if the QTL allele Q_{i2} is inherited;
- b_{i0} and b_{i1}^* are the random intercepts and random QTL effect in family i , for which we have $\mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1}^* \end{bmatrix} \sim \mathcal{N}(0, \Delta)$ in which $\Delta = \begin{bmatrix} \delta_{00} & \delta_{01} \\ \delta_{01} & \delta_{11} \end{bmatrix}$ with δ_{11}^* being the variance of QTL random effect;
- e_{ij} is the residual assumed to be *i.i.d.* distributed as $\mathcal{N}(0, \sigma_e^2)$.

Since the QTL alleles cannot be observed, x_{ij}^* is a missing value. A bull can have as many as hundreds of or even thousands of offspring. In such cases, the number of mixture components ($2^{\text{family size}}$) is a very huge number and the cost of computation (the memory and storage capacity required by the huge number of mixture components and the computing time, etc.) becomes prohibitively computationally expensive. If we let $z_{ij} = E(x_{ij}^* | \mathbf{M}_{obs})$ be the conditional expectation of x_{ij}^* given

the observed marker information \mathbf{M}_{obs} , the above model (4.3.1) is approximated into a linear mixed effects model (see details of model assumptions and derivations in Appendix B):

$$y_{ij} = \beta_0 + \beta_1 z_{ij} + b_{i0} + b_{i1} z_{ij} + e_{ij} \quad (4.3.2)$$

and in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}. \quad (4.3.3)$$

where \mathbf{Y} and \mathbf{e} are vectors of trait values and residuals, respectively; the fixed effects vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}'$ and the random effects $\mathbf{b} = \begin{bmatrix} \mathbf{b}'_1 & \mathbf{b}'_2 & \cdots & \mathbf{b}'_m \end{bmatrix}'_{2m \times 1}$, and the design matrices for fixed and random effects are

$$\mathbf{X} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \\ 1 & z_m \end{bmatrix}_{N \times 2}, \quad \mathbf{Z} = \begin{bmatrix} 1 & z_1 & & & \\ & & 1 & z_2 & \\ & & & \ddots & \\ & & & & 1 & z_m \end{bmatrix}_{N \times 2m}$$

with

$$\begin{aligned} \mathbf{b} &\sim \mathcal{N}(0, \Delta \otimes \mathbf{I}_m) \equiv \mathcal{N}(0, \text{diag}\{\underbrace{\Delta, \dots, \Delta}_m\}), \\ \mathbf{e} &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}). \end{aligned}$$

in which \otimes is the Kronecker product.

4.3.2 Hypothesis testing

Parameters of interest in QTL mapping are the QTL fixed effect parameter β_1 and the variance parameter of QTL random effect δ_{11} . Our modeling approach allows flexible hypotheses formulations testing the fixed QTL effects parameter β_1 or the random QTL effect variance parameter δ_{11} or both.

$$\begin{array}{llllll}
M_0: & y_{ij} & = & \beta_0 & + & \beta_1 z_{ij} & + & b_{i0} & + & b_{i1} z_{ij} & + & e_{ij} \\
M_1: & y_{ij} & = & \beta_0 & & & + & b_{i0} & + & b_{i1} z_{ij} & + & e_{ij} \\
M_2: & y_{ij} & = & \beta_0 & + & \beta_1 z_{ij} & + & b_{i0} & & & + & e_{ij} \\
M_3: & y_{ij} & = & \beta_0 & & & + & b_{i0} & & & + & e_{ij}
\end{array}$$

From the layout above, it is easily seen that there are alternative ways of constructing testing hypotheses. For example, the negative two times of the log likelihood ratio between M_1 (as the null model) and M_0 (the alternative model) can be used as the test statistic for testing on the fixed QTL effects only and the corresponding testing hypothesis is $H_{01} : \beta_1 = 0$ vs. $\beta_1 \neq 0$. Similarly the likelihood ratio test between M_2 and M_0 testing on the random QTL effect can be constructed under the hypothesis $H_{02} : \delta_{11} = 0$ vs. $\delta_{11} > 0$. So is the hypothesis testing both fixed and random QTL effects simultaneously: $H_{03} : \beta_1 = 0$ and $\delta_{11} = 0$ vs. $\beta_1 \neq 0$ or $\delta_{11} > 0$. For convenience in references, we denote the LRT statistic for hypothesis H_{xy} as LR_{xy} .

If we regard M_3 as the null model with respect to M_1 as the alternative model, that is, if we assume the QTL effects normally distributed with mean zero and variance δ_{11} , then M_1 closely matches the variance component (VC) model and the testing hypothesis, $H_{13} : \delta_{11} = 0$ vs. $\delta_{11} > 0$, corresponds to that of VC method. However, with respect to the question whether there is a QTL at the putative locus, we would like to reject the null hypothesis of no QTL if either of the two QTL parameters, β_1 and δ_{11} , are non-zero. It is instantly clear that the log likelihood ratio from H_{03} equals the sum of those from H_{01} and H_{13} , that is, $LR_{03} = LR_{01} + LR_{13}$.

For the purpose of chromosome scanning for QTL using our mixed model approach, the likelihood ratio test is performed at every putative position along the marker coverage region. Since both fixed and random effects are involved in testing, the maximum likelihood method is used for maximizing the likelihoods under the null and alternative hypotheses. The chromosome-wide significance threshold was obtained via 2500 within-family permutations CHURCHILL and DOERGE [1994]; DOERGE and CHURCHILL [1996]. These statistical testing analysis was performed using PROC MIXED in SAS v9.1.3.

4.3.3 Calculation of R^2

The best linear unbiased estimates (BLUEs) and best linear unbiased predictions (BLUPs) of the fixed and random effects parameters at the suggestive QTL can be obtained. The predicted trait value with respect to QTL was calculated as follows:

$$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 z_{ij} + \hat{b}_{i1} z_{ij} \quad (4.3.4)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are BLUEs of the fixed intercept and QTL effects and \hat{b}_{i1} is the BLUP of the QTL effect with respect to sire i , respectively. The random intercept effects have been pooled with the residuals. Then R^2 is computed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{i.})^2}. \quad (4.3.5)$$

4.4 Simulation Studies of Mixed Model Approach

Some simulation studies were performed to investigate some statistical aspects of our mixed model approach. There are many scenarios and we in particular were interested in two problems: (i) is the type I error under control? If the multiple families belonging to the same pedigree but are treated as if they are unrelated, will the type I error be impacted? (ii) In our mixed model approach, hypothesis testing on a putative QTL involves testing on both the fixed and random QTL effects. That is, the number of parameters being tested in the likelihood ratio test statistic is two, rather than one, for a single trait single QTL testing scenario. In such case, we were interested in knowing whether the 1.5 LOD-drop interval would still correspond to roughly 95% confidence level empirically?

4.4.1 Type I error for across-family analysis: pedigree versus unrelated

There are sufficient statistical theory and studies elucidating the point-wise variance component testing. However, it was of concern to us whether our approximation of pedigreed multiple families into unrelated families would impact the type I error rates and if so, in what way? To this end, we conducted a simulation-based comparison of type I errors as outlined below:

1. We adopted the number of families and family sizes from the experimental data of Dr. Ashwell's lab [MUNCIE *et al.*, 2006] (Table 3.4.1). The number of replicates was set to be 500.
2. There is a single chromosome of length 102 cM; markers are evenly spaced with 6cM apart; no QTL was assumed because we were only interested in the type I error.
3. We simulated the genotype data in two ways:
 - (a) Pedigree-based (PB): marker genotypes are simulated by abiding by the Mendelian rules and the pedigree relationships such that the paternal haplotypes in intermediary sires, who are both sons of a sire of previous generation and sires of his own family, would descend from his own sire with random crossing-over allowed.
 - (b) Unrelated (UN): Those families are regarded as if unrelated such that all sires' haplotypes are randomly sampled from the population.
 - (c) In either case, dam's gametes are always randomly sampled from the population.
 - (d) In either case, given a sire's haplotypes, the sire gamete to be transmitted to his offspring is subject to a stochastic process of whether, how many and where the crossing-over would occur. See detailed description of genotypic data simulation in Chapter 2 or in the following section of (4.4.2).

4. Phenotypic values of each individual are simulated using a null model without QTL: $y_{ij} = \beta_0 + e_{ij}$ with $\beta_0 = 10$ and $e_{ij} \sim \mathcal{N}(0, 1)$. The random intercept effect term has been omitted here because it is a nuisance parameter.
5. Mixed model analysis of both pedigree-based (PB) and unrelated (UN) simulation data were performed and the type I error rates were computed as the ratio between the number of replicates in which a significant signal is detected and the total number of replicates, *a.k.a.*, 500. We examined four alternative testing hypotheses: H_{01} , H_{02} , H_{03} and H_{13} , as specified above in section 4.3.2. Each replicate was permuted 1000 times and all permutation-based chromosome-wide maximum likelihood ratios are pooled across replicates to obtain the 90%, 95% and 99% percentiles as the respective significance thresholds.

The type I error rates are summarized in Table 4.4.1. It clearly shows that the type I errors are well under control in either simulation scenarios. For the data simulated as pedigree-based such that our mixed model's treatment of them as unrelated do not strictly hold, the estimated type I error rates approximate the nominal significance level quite well. If the data are simulated as unrelated, then our mixed model's treatment is right and the type I errors are also shown to be well under control. We recognized the fact that we used significance thresholds obtained as percentiles from a pool of thresholds across replicates and permutations. To that end, we examined the type I error rates if the replicate-specific permutation-based threshold was used and concluded that the differences were very minor (data not shown) and our observation that type I errors are under control still holds.

4.4.2 Choice of LOD-drop intervals

There has been simulation studies investigating the relationship between LOD-drop support intervals and confidence level [VISSCHER *et al.*, 1996; LAURIE *et al.*, 2008]. However, in those settings, the QTL effects being tested are fixed effects and only one effect parameter for a trait is being tested. The numerator degree of freedom of the point-wise test statistic is thus 1. However, in our mixed model framework,

Table 4.1: Type I error rates for treating pedigree-based families as unrelated. α is the significance level; PB: pedigree-based; UN: unrelated. H_{01} tests on QTL fixed effect β_0 only; H_{02} tests on QTL random effect δ_{11} only; H_{13} tests on both β_0 and δ_{11} simultaneously; H_{13} tests on QTL random effect δ_{11} under the assumption of $\beta_0 = 0$. Significance thresholds used for obtain these type I errors were the percentiles from the pooled chromosomal maximum likelihood ratio test statistics from 1000 permutations of each replicate. However, if for each replicate, the respective within-replicate permutation threshold is used, the type I error rates will only differ slightly from the table shown.

Method	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	PB	UN	PB	UN	PB	UN
H_{01}	0.0977	0.1253	0.0488	0.0601	0.0142	0.0092
H_{02}	0.0905	0.0966	0.0529	0.0488	0.0142	0.0122
H_{03}	0.0997	0.1058	0.0519	0.0610	0.0071	0.0102
H_{13}	0.0946	0.1038	0.0529	0.0498	0.0112	0.0173

when a QTL is being tested, both its fixed and random effects are under scrutiny. The number of parameters in the alternative model that are not present in the null model is two, at any putative locus. However, it is not known whether the 1.5 LOD-drop support interval would still correspond to roughly 95% confidence. To this end, we simulated some simple scenarios and evaluate the confidence level of variant LOD-drop intervals.

Simulation settings

There were in total 500 simulation replicates, each of which contains 50 unrelated paternal half-sib families with 40 individuals in each family. The choice of 40 for family size was a balance between sufficiently large family size such that the sire haplotypes will be reconstructed with almost certainty using our method described in Chapter 2, and the computational cost for the simulation data analysis.

A single chromosome of 70cM long was simulated with eight markers evenly spaced so that the inter-marker distance is 10cM apart. Every marker has four alleles with

allele frequencies randomly drawn from a uniform distribution then re-scaled to sum to unity. A single QTL was placed at 35cM with two QTL alleles of equal frequencies.

For each replicate, sires' haplotypes are randomly sampled according to the population allele frequencies at each marker, assuming marker linkage equilibrium. Thus, from one replicate to another, the number of sires and which sires that are segregating at QTL are random. For each individual, a Poisson random number, with the Poisson parameter equal to 0.7, is drawn to simulate the number of crossing-over in the respective sire. Given that one or more crossing-over has to take place, uniform random numbers are drawn to locate whether the crossing-over would take place and the respective two sire haplotypes will undergo crossing-overs at such locus (loci). One of the two thus-derived sire gametes will be randomly picked and assigned to the individual. The dam gamete for this individual is just randomly sampled according to the population allele frequencies.

Simulation of the phenotypic values was based on the following model:

$$y_{ij} = \beta_0 + b_{i0} + QTL + e_{ij} \quad (4.4.1)$$

where

- $\beta_0 = 5$, $b_{i0} \sim \mathcal{N}(0, 1)$ and $e_{ij} \sim \mathcal{N}(0, 1)$ for all $i = 1, \dots, m$ and $j = 1, \dots, n_i$;
- QTL effects are assumed additive only as $a, 0, -a$ for QTL genotypes QQ , Qq and qq , respectively; the variance due to QTL is then $V_{QTL} = 2p(1-p)a^2 = a^2/2$ where $p = 0.5$;
- QTL additive effect, a , was set as $a = \sigma_p/2$ so that the QTL variance accounts for $\sim 12.5\%$ of total phenotypic variance [COPPIETERS *et al.*, 1999].

QTL Analysis and power calculations

Due to computational concern, variance components analysis using SOLAR was not conducted. Instead, we compared the power and LOD intervals from least squares method and our mixed model method. We checked various LOD-drop intervals such

as 1, 1.5, 2.0, and 2.5. For each LOD-drop, the power of QTL mapping is defined as the ratio between the number of replicates in which the chromosomal maximum test statistic exceeds the threshold and its LOD-drop interval covers true QTL, and the number of total replicates (*a.k.a.*, 500). A closely related quantity called LOD-drop interval coverage probability was also calculated in a way similar to power but the denominator is the number of cases a maximum test statistic exceeds threshold. That is, the coverage probability is the ratio between power and the fraction of a positive signal detected. Lastly, the average width of LOD-drop intervals is calculated.

Results for LOD-drop intervals

The results of power, coverage probability and the width of LOD-drop intervals are summarized in Table 4.4.2. It seems that for such simple simulated scenarios, our mixed model has high coverage probabilities, that is, once some significant signal is detected, most likely the true QTL lies within the LOD-drop interval from the peak location. Also from Table 4.4.2, it seems that 1.5 LOD-drop interval would still be appropriate in that it offers good coverage probability with respect to the nominal 95% significance level. Another observation is that our mixed model approach achieves better statistical testing power than the least squares method, at least in these simulated scenarios. Note that in the least squares method, testing on a QTL involves testing on many QTL effects parameters jointly, one per family, thus the numerator degree of freedom of the test statistic is large, which may impact the testing power [SNEDECOR and COCHRAN, 1989]. In the contrast, in our mixed model formulation, for one QTL, there are at most two parameters needed to describe its effects.

Table 4.2: Confidence levels of various LOD-drop intervals from a simulation study. LS: least squares method; MX: mixed model method. Values in parentheses are standard errors of interval width.

Method	1.0 LOD-drop		1.5 LOD-drop		2.0 LOD-drop		2.5 LOD-drop	
	LS	MX	LS	MX	LS	MX	LS	MX
Power (%)	70.4	83.4	80.2	87.4	85.4	88.8	88.8	88.8
Coverage (%)	74.9	93.9	85.3	98.4	90.9	100	94.5	100
Width (cM)	15.4	26.9	20.4	36.6	25.0	45.5	29.4	52.5
(se)	(0.76)	(1.45)	(0.98)	(1.79)	(1.11)	(1.92)	(1.27)	(1.86)

4.5 Application to Experimental Data

4.5.1 Our mixed model analysis

It is common with dairy cattle that families may be related and such genetic relatedness can be quite complicated. Take our experimental pedigree as an example: there are 13 families belonging to three generations [MUNCIE *et al.*, 2006]. The relationships between the founder sires involve various types such as the father-son relationship, sib-ship, cousin-ship, uncle-nephew relationship and more complicated relationships. It could be very challenging and computationally infeasible to properly account for such relationships among sires. On the other hand, in the half-sib design, dams were not genotyped at all and thus their genetic contributions to the offspring are unknown. Such maternal uncertainty overshadows the relatedness between sires. What's more, to peel the pedigree, further assumptions on marker allele frequencies and marker linkage equilibrium in dam populations are needed that may not even be valid. Therefore, we opted to treat the half-sib families as unrelated to balance between genetic information and computation costs.

4.5.2 Variance components analysis

Variance components analysis can be separated from IBD computation; that is, given an appropriately calculated and formatted data file containing IBD coefficients,

SOLAR is capable of performing REML analysis of the variance components models. We approached the calculation of IBD values by both the SOLAR way and an approximation way of considering only the paternally IBD relationships. We denote the IBD coefficients computed by SOLAR as IBD_{solar} and the IBD coefficients by our approximations as IBD_{pat} .

IBD computation using SOLAR

For the purpose of variance components analysis, IBD matrices at genome scanning positions need to be calculated. However, this turned out to be a challenging task for the experimental data. In our initial attempts to compute the IBD coefficients, we treated the whole multi-generational pedigree as a whole. We tried in vain three most popular genetics software programs: Merlin [ABECASIS *et al.*, 2002], Loki [HEATH, 1997] and SOLAR [ALMASY and BLANGERO, 1998]. Merlin implements a branch-and-conquer algorithm but has a 2 GB maximum limit for memory allocation and directly failed for our data. Loki implements a Monte Carlo Markov Chain algorithm for computing IBD. However, outputs from independent MCMCs from different computers and random seeds do not even seem to converge after month-long computation. SOLAR could not yield IBD results for even a single marker after month-long computation, either.

Facing this computational difficulty and due to the limited computation resources available, we opted to treat the multiple families as unrelated, although they belong to the same pedigree and are, in fact, related. Even after such simplification of relationships, Merlin and Loki still failed and it takes SOLAR more than one day to compute the IBD coefficients for a single marker. We took advantage of the available computing resources and employed SOLAR for IBD computations of different markers simultaneously, which significantly shortened the whole time period of IBD calculation. Once the marker IBD coefficients were obtained, we computed the multipoint IBD coefficients for every 1cM position along the chromosomal segment.

IBD computation considering paternal alleles only

Half-sibship between offspring from the same parent say a sire offers an opportunity to use the following simplified formula for computing multipoint IBD coefficients between the paternally derived alleles only in offspring j and j' of the same sire i :

$$\text{IBD}(j, j') = \frac{\Pr(Q_{ij}^s | \mathbf{M}_{obs}) \Pr(Q_{ij'}^s | \mathbf{M}_{obs}) + \Pr(q_{ij}^s | \mathbf{M}_{obs}) \Pr(q_{ij'}^s | \mathbf{M}_{obs})}{4} \quad (4.5.1)$$

where Q_{ij}^s or q_{ij}^s denote the paternally derived QTL alleles, Q_i and q_i , in offspring j and j' , respectively. The upper and lower cases denote the two allele variants in the sire i . Only when the two paternal alleles are of identical case, there will be non-zero paternal IBD probabilities between these two offspring. Obviously, the IBD value is 0 between unrelated two individuals and 1 between one individual and oneself. The IBD thus calculated is not the same as those from SOLAR because SOLAR's IBD computation uses the marker genotypes and accounts for both paternal and maternal IBD relationships whenever possible.

4.5.3 Mixed model analysis results

Our mixed model analysis detected suggestive QTL for four reproductivity-related traits: DPR, PL, SCS and PDB but no suggestive QTL for traits of calving ease (CE) and milk composition traits (milk yield, fat and protein yield or percentage).

Table 4.5.3 summarizes the detected QTL from our across-family analysis and Figure 4.1 shows their LOD score profiles with the support intervals and peak locations. A suggestive QTL affecting DPR is placed at 49.6cM with a 1.5 LOD-drop support interval ranging from 25.6 to 64.6cM. A suggestive QTL affecting PDB was detected at 70.6cM with a 1.5 LOD-drop interval ranging from 30.6 to 83.6cM. A suggestive QTL affecting SCS, an indirect measure of mastitis, was detected at 36.6cM with a 1.5 LOD-drop interval ranging from 24.6 to 48.6cM. There are two suggestive QTL affecting PL, the period an animal is active in the milking herd before culling – one at 35.6cM and another at 79.6cM, with their 1.5 LOD-drop intervals 24.6 - 63.6cM

Table 4.3: Across-family mixed model analysis results for QTL mapping.

Trait	Max LOD	Threshold (95%)	Peak Location (cM)	1.5 LOD-drop Support Interval (cM)	QTL Effects Estimates		R^2 (%)
					Mean	Variance	
PTA DPR	1.9	1.6	50	26 - 65	-0.07	0.05	2.3
PTA SCS	2.3	1.8	37	25 - 49	0.03	0.002	2.5
PDB	1.8	1.8	71	31 - 84	-0.33	0.10	0.9
PTA PL	2.8	1.7	36	25 - 64	-0.04	0.17	3.6
DD PL	1.9	1.8	80	25 - 84	0.47	< 0.0001	0.6

and 24.6 - 83.6cM, respectively.

One important feature of our across-family mixed model analysis is that we can decompose the sources of signal by formulating different testing hypotheses. To illustrate the different behaviors of the likelihood ratio profiles, we present Figure 4.2 in which there are four curves of the log likelihood ratios for each of the traits with suggestive QTL detected.

It is easily observed that there are different patterns of the sources of signal for different traits. For traits PTA DPR and PTA SCS, most of the signal is from the random QTL effect that depicts the variation of QTL effects across families and the log likelihood ratios for testing on the fixed QTL effect is essentially flat at zero. However, for the trait PDB, the log likelihood ratios for testing on the fixed QTL effect were larger than that for testing on random QTL effects across the region. At 71cM, the peak location corresponding to the hypothesis (LR_{03}) that tests on the overall QTL effects, both hypothesis testings of H_{01} and H_{13} contribute equally to the overall signal strength. Yet it is a different case for the trait DD PL: in the first half of the chromosome region, there is a strong indication of a QTL at ~ 36 cM that exhibits solely the random effects; whereas there is another peak towards the end of the region around 80cM that is suggestive of another QTL exhibiting mainly fixed effects. The plot for PTA PL shows a similar pattern as DD PL, but the log likelihood ratios for testing on the QTL fixed effects towards the end of the region did not yield as high as that of DD PL.

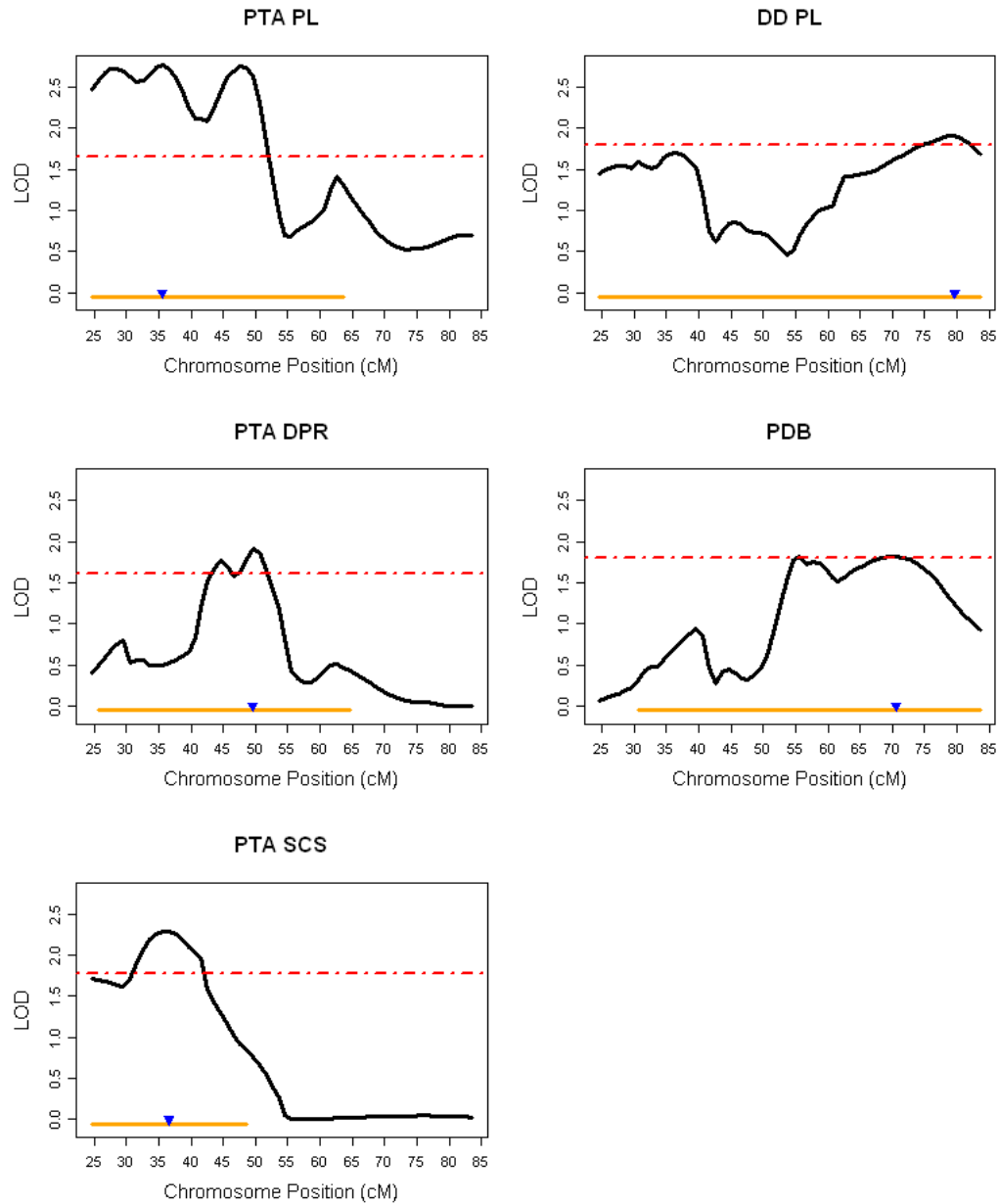


Figure 4.1: Across-family analysis results. The 95% threshold values were obtained from 2500 permutations. Traits are predicted transmitting ability (PTA) or daughter deviation (DD) for daughter pregnancy rate (DPR), productive life (PL), somatic cell score (SCS) and percent difficult births (PDB).

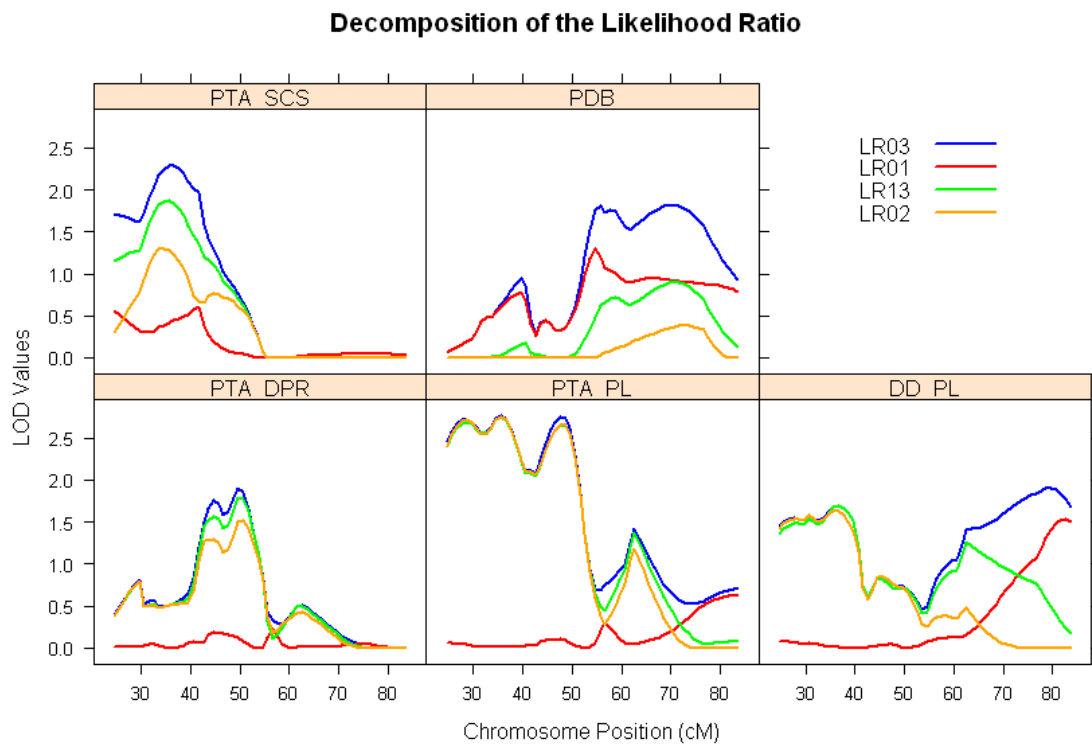


Figure 4.2: Decomposition of likelihood ratios in our mixed model analysis. LR_{xy} is the log likelihood ratio between models M_x and M_y . Traits are predicted transmitting ability (PTA) or daughter deviation (DD) for daughter pregnancy rate (DPR), productive life (PL), somatic cell score (SCS) and percent difficult births (PDB).

4.5.4 Comparison with variance components analysis results

In Figure 4.3, LRT profiles of testing on QTL random effects only by using the three methods are overlaid: the LR_{03} and LR_{13} from our mixed model, the LRT from SOLAR using IBDpat ($LR_{ibdp\text{at}}$) and the LRT from SOLAR by using SOLAR's own IBD (LR_{solar}).

We compared our mixed model method with the variance component method implemented in SOLAR software for single QTL modeling. The hypothesis H_{13} in our mixed model method is similar to QTL variance component testing on QTL effects in VC analysis because both hypotheses test on QTL effects' variance under the assumption that QTL effects are normally distributed with mean zero among families. The VC analysis in SOLAR failed to detect the suggestive QTL for SCS and PDB, and gave slightly different peak locations for the trait DPR. The results for the trait PL were interesting: our mixed model results suggested two genetic factors affecting the trait – one in the range of 24 - 40cM and another in the range of 70 - 80cM (manifested by the H_{13} and H_{01} curves, respectively; see Figure 4.2). However, the VC method only captures the first factor that is primarily due to random effects (Figure 4.3).

On the other hand, the LRT profile of variance components analysis using IBDpat computed is highly similar with that of LR_{13} (Figure 4.3), suggesting that our mixed model is able to capture most signals that variance components can. However, such similarity is not well conserved between LR_{13} and LR_{solar} (QTL random effect testing using IBD calculated by SOLAR), which could be due to the fact that maternal marker information was used by SOLAR.

4.6 Discussions

QTL mapping is more complicated in multiple families for various reasons. For example, a QTL may not be heterozygous in all families. In families where the QTL is heterozygous, there may be different pairs of QTL alleles segregating in different families, which is a reasonable assumption for most outbred populations. Even if

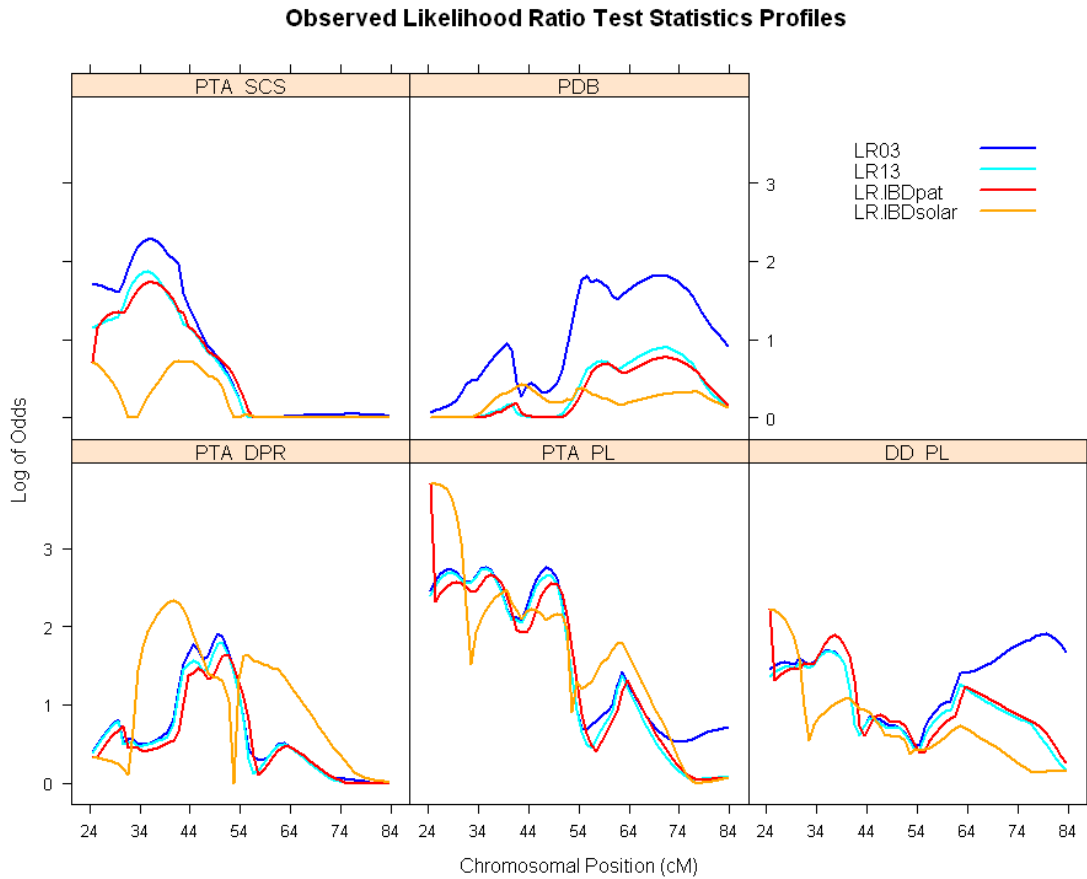


Figure 4.3: Comparison of likelihood ratio test profiles on QTL random effects only. LR_{xy} is the log likelihood ratio between models M_x and M_y ; LR.IBDpat and LR.IBDsolar are log likelihood ratios testing on QTL variance components with the IBD coefficients computed by the approximation method using paternal information only or the multipoint IBD calculation method in SOLAR. Traits are predicted transmitting ability (PTA) or daughter deviation (DD) for daughter pregnancy rate (DPR), productive life (PL), somatic cell score (SCS) and percent difficult births (PDB).

the same pair of QTL alleles are segregating in multiple families, the directions of QTL effects (promoting or demoting phenotypic performance) may be random across families. Since a QTL will escape detection in single family analysis if it is not segregating in that particular family, analysis of data from multiple families would help to address this potential problem.

In our across-family mixed model approach, half-sib families were assumed to be unrelated. Although this assumption was not strictly correct with regard to experimental data, our approach was appropriate for computational feasibility and its practical balance between the inter-family paternal relatedness and the genetic uncertainty caused by un-genotyped dams. Simulation studies based on the same pedigree relationship and the sample sizes as that found in the experimental data showed that our mixed model approach has the type I error well under control (Table 4.4.1).

Our mixed effects approach tries to dissect the QTL effects into the fixed effects (across-family mean) and random effects (between-family variation). This approach clearly is advantageous in handling QTL with multiple alleles and varied QTL segregation modes across many families. Statistically, it only takes two parameters, the across-family mean effect (β_1) and the between-family variation (δ_{11}) to describe effects of a QTL. In contrast, the least squares method needs one QTL effect parameter per family, which results in large numerator degrees of freedom in the hypothesis testing that is detrimental to statistical testing power [SNEDECOR and COCHRAN, 1989]. This is a concern for LS analysis involving a large number of families. Because our experimental data only contains a moderate number of families, such concern would not become a serious issue and except for the trait PL, LS analysis gives generally similar LRT profiles as well as peak locations (data not shown).

On the other hand, the variance components method such as the one implemented in SOLAR [ALMASY and BLANGERO, 1998], in which the QTL effects are assumed to be normally distributed with mean zero, has also been widely applied in human genetics [ALMASY and BLANGERO, 1998] and animal breeding [GEORGE *et al.*, 2000]. One of its advantages is that it uses the IBD and kinship coefficients to capture the inter-individual relatedness across a pedigree, instead of being limited within each family. However, such feature is not applicable to our experimental data analysis

simply because of prohibitive computational costs for obtaining IBD coefficients due to the very large family sizes if families within the same pedigree are to be considered related. In the end, we opted to treat families as unrelated and computed the IBD coefficients, which was still a time-consuming task.

The variance component analysis results show slightly different LOD curves from our mixed model results for the DPR trait, which is expected since the two statistical models were specified slightly differently and the variance-covariance matrix of the random effects were also numerically different. The VC analysis available in SOLAR failed to detect suggestive QTL that our mixed model method detected for traits SCS and PDB, gave slightly different peak location of the LRT curve for the trait DPR, and only captured random-effects related signals while our mixed model also detected fixed-effects-related signals, as demonstrated in the case of the trait PL. Above being said, if we adopt an IBD calculated only with respect to paternal allele transmission, then we get highly similar LRT profile curves between LR_{ibdpat} from the variance component analysis in SOLAR and LR_{13} from our mixed model analysis.

Chapter 5

Summary and Discussions

5.1 Summary and Discussions

In this dissertation, we have made endeavors to develop or improve methods that are part of the whole QTL analysis flow. In the initial stage of analyzing outbred populations, the marker linkage phases in parental individuals/lines need to be inferred. Due to the usual large number of progeny for dairy cattle half-sib design, such inference generally leads to highly accurate reconstructed haplotypes in sires. However, existing methods might be highly affected by missing data, small family sizes and non-negligible recombination. To this end, our proposed method of inferring sire haplotypes proves to overcome these aforementioned challenges and consistently achieves great accuracy in reconstructed haplotypes. Future extension of our method to general sibships is straightforward. Two problems that are more interesting are (i) to improve speed performance and (ii) inferring most probable haplotypes in offspring. For the former task, we propose to use a deterministic non-parametric index to rapidly fish out the most probable haplotype configurations such that the exact likelihood computation on such a subset of highly probable haplotype configurations would only need a relative small amount of computation resources.

For the latter task, there have been suggested many choices such as Gibbs sampling [MEUWISSEN *et al.*, 2002] or other types of MCMC [LEE *et al.*, 2005] to emulate the genetic inheritance based on Mendelian rules. However, we would like to eye on

a straightforward deterministic method of assigning parental gametes into an offspring. The difficulty of deducing parental gametes in an offspring is mainly in the following scenarios, given parental haplotypes already known: if it is determined that one crossing-over has to take place between two marker loci and if there are one or more other un-informative markers situated in-between these two markers, choice of crossing-over locations would lead to different deduced parental gametes for the offspring. Taking a frequentist's approach, we would tend to pick the mid-point of the interval as the crossing-over location. If there are markers very close to the mid-point, then alternative locations of crossing-over with respect to these markers will be considered and either the most likely or all these close to equally likely parental gametes will be considered with respective relative weight carried over for future usage such as IBD calculations.

For the purpose of QTL mapping in large single dairy cattle half-sib family, once the sire haplotypes are inferred, the subsequent QTL analysis follows. To this respect, we adapted the multiple interval mapping (MIM) framework to this outbred design and through a case example, we showed that MIM approach offers more powerful testing for detecting QTL and better resolutions in determining support intervals for the QTL location. Not only theoretical derivations of our extensions such as the modeling of heteroscedastic errors were carried and implemented into our in-house software – Windows QTL Cartographer, revisions to the data file format for WinQTLCart were also introduced with the help of Dr. Shengchu Wang.

Lastly, we have addressed the task of jointly analyzing QTL across families, in the hope of both overcoming the uncertain segregation status of one or several QTL in any single family, and improving statistical analysis power by the enlarged sample size and joint utilization of the information. Our mixed model approach is shown to be more powerful than variance components and flexible in constructing alternative testing hypotheses. A simple case simulation comparison was also suggestive of higher power of our mixed model than the least squares method.

From our own analysis of within-family and across-family, we noted the following: due to artificial insemination technology, a dairy cattle founder sire may have as many as hundreds of sons, as is in our experimental data. This is distinctly different

from human populations, for which the variance components method was originally proposed. Should QTL be segregating in such large half-sib families, the within-family analysis would yield good power and precision in mapping quantitative trait loci. On the other hand, given a suggestive QTL in one family, inclusion of other families in which this particular locus is non-segregating into an across-family joint analysis would only lead to an elevated significance threshold. If the suggestive QTL only has weak or mild effects in the original single family, it would be very likely that the across-family analysis would not detect it as significant. This was exemplified by our results of within- and across-family analyses (Tables 3.4.2 and 4.5.3), where some suggestive QTL were detected in within-family analysis but not the across-family analysis. Furthermore, the within-family analysis gives better estimation about the QTL (location, support interval and effect size), whereas the across-family analysis offers validating information for the existence of QTL across the families and the overall heritability. For instance, the estimated R^2 for PTA DPR trait was 2.3% in Table 4.5.3 while its heritability was previously reported as 4% [VANRADEN *et al.*, 2004]. Data from multiple dairy cattle families are still needed for mapping quantitative trait loci in order to overcome the uncertain segregation status at putative loci within a single family. It is preferable that these families be large and informative, which can be partly achieved through selective genotyping and preliminary analysis of a reduced subset of markers.

5.2 Towards Fine Mapping

All we have discussed reside mainly within the domain of linkage analysis. However, demands to precisely pinpointing QTL within a less than 1cM segment are strong so that the positional cloning would be practically carried on. Linkage information comes from the recombination within the recorded pedigrees, thus the precision would not be as high as so desired. Increasing marker density, however, would not greatly improve linkage analysis precision, either, due to the limited number of observable recombination within the recorded pedigree.

5.2.1 Linkage disequilibrium linkage analysis - LDLA

An increasingly promising direction is to jointly make use of the linkage and linkage disequilibrium information for the purpose of mapping QTL. Linkage disequilibrium reflects historical recombination events over many generations ever since the mutation or admixture events took place. FARNIR *et al.* [2000] showed that there are residual LD in the bovine population for up to 2cM. Therefore, under appropriate assumption of the ancient QTL mutation/admixing events and a coalescent history of the population, the LD-based IBD coefficients may be calculated between haplotypes from the population. With respect to dairy cattle half-sib design, sire haplotypes and dam gametes would be suitable for such computation and are called base haplotypes. However, the paternal gametes that are inferred in the half-sib offspring are not base haplotypes, because they are descending from their sires' haplotypes via linkage with or without crossing-over during gametogenesis.

MEUWISSEN *et al.* [2002] proposed a three-step approach of LDLA as outlined below:

1. Deduce and infer haplotypes in sires, paternal and maternal gametes in offspring. The sire haplotypes and maternal gametes are regarded as base haplotypes and paternal gametes, however, non-base haplotypes.
2. Compute multipoint IBD coefficients between base haplotypes under the assumption that the putative QTL can be traced back to a single ancestral mutation and the LD-based IBD reflects the QTL segregation in the current recorded pedigree. A set of formulae were derived based on the coalescence theory and exhaustive enumerations of possible recombination/IBD scenarios between the putative locus and marker-haplotypes [MEUWISSEN and GODDARD, 2001]. The IBD between base and non-base haplotypes and between non-base haplotypes can then be computed by using both LD and linkage information [MEUWISSEN *et al.*, 2002]. The above strategy of IBD computation can also be represented in Table 2.

Table 5.1: Haplotype-based IBD coefficient matrix at a locus p . [a] refers to LD-based IBD calculation between base haplotypes and MEUWISSEN and GODDARD [2001] derived formulas based on coalescence theory, [b] refers to IBD computation between base and non-base haplotypes in a way as MEUWISSEN *et al.* [2002] did, and [c] refers to IBD between sire gametes within the pedigree that is to be computed in the last because its calculation would require IBD computation in blocks of [b] having been completed.

	Sire Haplotypes	Dam Gametes	Sire Gametes
Sire Haplotypes	[a]	[a]	[b]
Dam Gametes	[a]	[a]	[b]
Sire Gametes	[b]	[b]	[c]

3. A haplotype-based variance component model of analyzing QTL variations at a putative locus is to be evaluated in a chromosomal scanning manner:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_h\mathbf{h} + \mathbf{Z}_u\mathbf{u} + \mathbf{e} \quad (5.2.1)$$

where

- \mathbf{y} , $\boldsymbol{\beta}$, \mathbf{u} , \mathbf{e} are vectors of phenotypic data, fixed effects, polygenic effects and residuals, respectively;
- \mathbf{X} is the incidence matrix relating fixed effects to individuals;
- \mathbf{h} is a vector of haplotype effects, comprised of both base and non-base haplotypes;
- \mathbf{Z}_h is the incidence matrix relating individuals with the pertaining haplotypes;
- $\mathbf{h} \sim \mathcal{N}(0, \sigma_Q^2 \mathbf{G})$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{A})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$.

Maximum likelihood analysis of such a model is usually fulfilled by using REML and likelihood ratio test is the usual device for testing QTL. The old problem of finding appropriate significance threshold remains but only more challenging: the marginal distribution of the point-wise likelihood ratio test statistic on a variance

component (σ_Q^2) now follows a half-half mixture of χ_0^2 and χ_1^2 but the distribution of the chromosomal-wide maximum likelihood ratio is not known. The permutation procedure by CHURCHILL and DOERGE [1994]; DOERGE and CHURCHILL [1996] is not readily applicable, because the additive polygenic effects need to be preserved while QTL effects are to be permuted, which still does not seem to be possible for a general pedigree yet.

5.2.2 Variant methods and prospective extensions

The distinct feature of the LDLA approach as described above is the fact that the IBD coefficients among haplotypes have the LD information incorporated along with the linkage information. Applications of such method have been reported for fine mapping QTL into less than 1Mb segment or specific SNPs [SCHNABEL *et al.*, 2005; OLSEN *et al.*, 2008; KIM *et al.*, 2009]. Other variants have also been reported: LEE *et al.* [2005] replaces the Gibbs sampler with a reversible jump MCMC for the haplotyping step; BLOTT *et al.* [2003] and DRUET *et al.* [2008] used the hierarchical clustering technique of the base haplotypes for obtaining a dimension-reduced vector of haplotypes for a variance component model similarly specified as in equation (5.2.1).

All of these method have been treating the maternal gametes as if the haplotypes in dams, which is not strictly true. Although we do not get to observe the genotypes in dams, it would be reasonable to assume that a portion of the maternal gametes, at a similar fraction with that of sire haplotypes, are actually recombinants from the respective maternal haplotypes. It is sensible to assume that a pool of maternal haplotypes (fewer than the maternal gametes) would better convey the LD information from the ancestral mutation to the current pedigree. The clustering of base haplotypes is one way of retrieving a compact set of distinct haplotypes. Another possible way is to apply some population-based haplotyping method to the large number of maternal gametes and deduce the maternal haplotypes. Then the IBD computation and variance component modeling can be conducted using these compact set of base haplotypes, etc.

Chromosome-wide significance threshold

Although an approximated χ_1^2 distribution for the chromosomal-wide maximum likelihood ratio test statistic has been suggested by GEORGE *et al.* [2000], it may still be worth studying how to obtain a chromosomal-wide significance threshold through some statistical procedure, since such a threshold is a function of the chromosomal length and sample sizes, etc. Score tests on variance components have been applied to haplotype-based association mapping [TZENG and ZHANG, 2007]. Therefore, the score-based re-sampling strategy similar to those in ZOU *et al.* [2004] and LAURIE *et al.* [2008] would seem to be a likely solution.

Acknowledgement

The author thanks Fang Hu for helpful discussions about fine mapping methodology and the kind offering of the trial version of fine mapping programs.

Bibliography

- ABECASIS, G. R., S. S. CHERNY, W. O. COOKSON, and L. R. CARDON, 2002
Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**: 97–101.
- ALMASY, L., and J. BLANGERO, 1998 Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**: 1198–1211.
- AMOS, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet.* **54**: 535–543.
- ASHWELL, M. S., D. W. HEYEN, T. S. SONSTEGARD, C. P. TASSELL, Y. DA, *et al.*, 2004 Detection of quantitative trait loci affecting milk production, health, and reproductive traits in holstein cattle. *Journal of dairy science* **87**: 468–475.
- BARUCH, E., J. I. WELLER, M. COHEN-ZINDER, M. RON, and E. SEROUSSI, 2006 Efficient inference of haplotypes from genotypes on a large animal pedigree. *Genetics* **172**: 1757–1765.
- BLOTT, S., J.-J. KIM, S. MOISIO, A. SCHMIDT-KUNTZEL, A. CORNET, *et al.*, 2003 Molecular Dissection of a Quantitative Trait Locus: A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated With a Major Effect on Milk Yield and Composition. *Genetics* **163**: 253–266.
- BOEHNKE, M., 1991 Allele frequency estimation from data on relatives. *Am J Hum Genet.* **48**: 22–25.

- BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 1572–1577.
- CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- COPPIETERS, W., A. KVASZ, J. J. ARRANZ, and B. GRUISART, 1999 The great-grand-daughter design: a simple strategy to increase the. *Gene.Res.Camb.* **74**: 189–199.
- COPPIETERS, W., A. KVASZ, F. FARNIR, J.-J. ARRANZ, B. GRISART, *et al.*, 1998 A Rank-Based Nonparametric Method for Mapping Quantitative Trait Loci in Outbred Half-Sib Pedigrees: Application to Milk Production in a Granddaughter Design. *Genetics* **149**: 1547–1555.
- CREPIEUX, S., C. LEBRETON, B. SERVIN, and G. CHARMET, 2004 Quantitative Trait Loci (QTL) Detection in Multicross Inbred Designs: Recovering QTL Identical-by-Descent Status Information From Marker Data. *Genetics* **168**: 1737–1749.
- CURTIS, D., and P. C. SHAM, 1995 Using risk calculation to implement an extended relative pair analysis. *Annals of Human Genetics* **58**: 151–162.
- DAVIS, S., M. SCHROEDER, L. GOLDIN, and D. WEEKS, 1996 Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *American Journal of Human Genetics* : 867–880.
- DING, X., Q. ZHANG, C. FLURY, and H. SIMIANER, 2006 Haplotype reconstruction and estimation of haplotype frequencies from nuclear families with only one parent available. *Hum Hered* **62**: 12–19.
- DOERGE, R. W., and G. A. CHURCHILL, 1996 Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285–294.

- DOERGE, R. W., Z. B. ZENG, and B. S. WEIR, 1997 Statistical issues in the search for genes affecting quantitative. *Statistical Science* **12**: 195–219.
- DRUET, T., S. FRITZ, M. BOUSSAHA, S. BEN-JEMAA, F. GUILLAUME, *et al.*, 2008 Fine Mapping of Quantitative Trait Loci Affecting Female Fertility in Dairy Cattle on BTA03 Using a Dense Single-Nucleotide Polymorphism Map. *Genetics* **178**: 2227–2235.
- DU, F.-X., B. W. WOODWARD, and S. K. DENISE, 1998 Haplotype Construction of Sires with Progeny Genotypes Based on an Exact Likelihood. *J. Dairy Sci.* **81**: 1462–1468.
- ELSTON, R. C., and J. STEWART, 1971 A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**: 523–542.
- FALCONER, D. S., and T. F. MACKAY, 1996 *Introduction to Quantitative Genetics*. Prentice Hall.
- FARNIR, F., W. COPPIETERS, J.-J. ARRANZ, P. BERZI, N. CAMBISANO, *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**: 220–227.
- FERNANDO, R. L., C. STRICKER, and R. C. ELSTON, 1994 The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance. *Theoretical and Applied Genetics* **88**: 573–580.
- FULKER, D., S. CHERNY, and L. CARDON, 1995 Multipoint interval mapping of quantitative trait loci using sib pairs. *American Journal of Human Genetics* **56**: 1224–1233.
- GAO, G., and I. HOESCHELE, 2008 A rapid conditional enumeration haplotyping method in pedigrees. *Genet Sel Evol.* **40**: 25–36.
- GAO, G., I. HOESCHELE, P. SORENSSEN, and F. DU, 2004 Conditional Probability Methods for Haplotyping in Pedigrees. *Genetics* **167**: 2055–2065.

- GEMAN, S., and D. GEMAN, 2000 Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.
- GEORGE, A. W., P. M. VISSCHER, and C. S. HALEY, 2000 Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics* **156**: 2081–2092.
- GEORGES, M., 2007 Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annual Review of Genomics and Human Genetics* **8**: 131–162.
- GEORGES, M., D. NIELSEN, M. MACKINNON, A., and MISHRA, 1995 Mapping quantitative trait loci controlling milk production in dairy. *Genetics* **139**: 907–920.
- GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
- GRIGNOLA, F. E., I. HOESCHELE, and B. TIER, 1996 Mapping quantitative trait loci in outcross populations via residual maximum likelihood. i. methodology. *Genetics selection evolution* **28**: 479–490.
- HALDANE, J., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of genetics* **8**: 299–309.
- HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.
- HALEY, C. S., S. A. KNOTT, and J.-M. ELSSEN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**: 1195–1207.
- HASTINGS, W. K., 1970 Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**: 97–109.
- HEATH, S. C., 1997 Markov chain monte carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.

- HOESCHELE, I., 1988 Statistical techniques for detection of major genes in animal breeding data. *Theoretical and Applied Genetics* **76**: 311–319.
- HOESCHELE, I., 2007 *Handbook of Statistical Genetics*, chapter Mapping quantitative trait loci in outbred pedigrees. Wiley, 3rd edition, 623–677.
- HOESCHELE, I., P. UIMARI, F. E. GRIGNOLA, Q. ZHANG, and K. M. GAGE, 1997 Advances in Statistical Methods to Map Quantitative Trait Loci in Outbred Populations. *Genetics* **147**: 1445–1457.
- INDURY, R., and R. ELSTON, 1997 A faster and more general hidden marko model algorithm for multipoint likelihood calculations. *Nature Genetics* **47**: 197–202.
- JANSEN, R. C., 1993 Interval Mapping of Multiple Quantitative Trait Loci. *Genetics* **135**: 205–211.
- JIANG, C., and Z.-B. ZENG, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.
- JUDSON, R., and J. C. STEPHENS, 2001 Notes from the snp vs. haplotype front. *Pharmacogenomics* **2**: 7–10.
- KAO, C. H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**: 855–865.
- KAO, C. H., Z.-B. ZENG, and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
- KIM, E.-S., P. J. BERGER, and B. W. KIRKPATRICK, 2009 Genome-wide scan for bovine twinning rate qtl using linkage disequilibrium. *Animal Genetics* .
- KNOTT, S., C. HALEY, and R. THOMPSON, 1991 Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* **68**: 299–311.
- KNOTT, S. A., J. M. ELSSEN, and C. S. HALEY, 1996 Methods for multiple-marker mapping of quantitative trait loci in. *Theor Appl Genet* **93**: 71–80.

- KRUGLYAK, L., M. J. DALY, M. P. REEVE-DALY, and E. S. LANDER, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet.* **58**: 1347–1363.
- KRUGLYAK, L., and E. S. LANDER, 1995 A Nonparametric Approach for Mapping Quantitative Trait Loci. *Genetics* **139**: 1421–1428.
- LANDER, E., and P. GREEN, 1987 Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**: 2363–2367.
- LANDER, E. S., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using rflp. *Genetics* **121**: 185–199.
- LANGE, K., and M. BOEHNKE, 1983 Extensions to pedigree analysis v. optimal calculation of mendelian likelihoods. *Hum Hered* **33**: 291–301.
- LAURIE, C., S. WANG, L. A. CARLINI-GARCIA, and Z. B. ZENG, 2008 Submitted.
- LEE, S. H., J. H. J. VAN DER WERF, and B. TIER, 2005 Combining the Meiosis Gibbs Sampler With the Random Walk Approach for Linkage and Association Studies With a General Complex Pedigree and Multimarker Loci. *Genetics* **171**: 2063–2072.
- LI, J., and T. JIANG, 2003 Efficient inference of haplotypes from genotypes on a pedigree. *J Bioinfo Comp Biol* **1**: 2003.
- LI, J., and T. JIANG, 2004 An exact solution for finding minimum recombinant haplotype configurations on pedigrees with missing data by integer linear programming. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. ACM, New York, NY, USA, 20–29.
- LIN, S., and T. P. SPEED, 1997 An algorithm for haplotype analysis. *J Comput Biol.* **4**: 535–546.

- M., V. P., H. C. S., and K. S. A., 1996 Mapping qtls for binary traits in backcross and f2 populations. *Genetical Research* **68**: 55–63.
- MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**: 480–488.
- METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER, 1953 Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092.
- MEUWISSEN, T. H., and M. E. GODDARD, 2001 Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution* **33**: 605–634.
- MEUWISSEN, T. H. E., A. KARLSEN, S. LIEN, I. OLSAKER, and M. E. GODDARD, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**: 373–379.
- MORTON, N., and C. MACLEAN, 1974 Analysis of family resemblance iii: complex segregation analysis of quantitative traits. *American Journal of Human Genetics* **26**: 489–503.
- MUNCIE, S. A., J. P. CASSADY, and M. S. ASHWELL, 2006 Refinement of quantitative trait loci on bovine chromosome 18 affecting. *Animal Genetics* **37**: 273–275.
- NEWTON, M. A., and A. E. RAFTERY, 1994 Approximate bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **56**: 3–48.
- O’CONNELL, J. R., 2000 Zero-recombinant haplotyping: Applications to fine mapping using snps. *Genetic Epidemiology* **19**: S64–S70.
- OLSEN, H. G., T. H. E. MEUWISSEN, H. NILSEN, M. SVENDSEN, and S. LIEN, 2008 Fine Mapping of Quantitative Trait Loci on Bovine Chromosome 6 Affecting Calving Difficulty. *J. Dairy Sci.* **91**: 4312–4322.

- QIAN, D. J., and L. BECKMANN, 2002 Minimum-recombinant haplotyping in pedigrees. *American Journal of Human Genetics* **70**: 1434–1445.
- SATAGOPAN, J., and B. YANDELL, 1996 Estimating the number of quantitative trait loci via bayesian model determination. In *Contributed Paper, Session on Genetic Analysis of Quantitative Traits and Complex Diseases*. Joint Statistical Meetings, Chicago, IL.
- SCHNABEL, R. D., T. S. SONSTEGARD, J. F. TAYLOR, and M. S. ASHWELL, 2005 Whole-genome scan to detect qtl for milk production, conformation, fertility and functional traits in two us holstein families. *Animal Genetics* **36**: 408–416.
- SEATON, G., C. S. HALEY, S. A. KNOTT, M. KEARSEY, and P. M. VISSCHER, 2002 Qtl express: mapping quantitative trait loci in simple and complex. *Bioinformatics* **18**: 339–340.
- SELF, S. G., and K. Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood. *JASA* **82**: 605–610.
- SIEBERTS, S., and E. SCHADT, 2007 *Handbook of Statistical Genetics*, chapter Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci. Wiley, 3rd edition, 296–326.
- SILVA, L., 2009 *Multiple Trait Multiple Quantitative Trait Loci Inferences*. Ph.D. thesis, North Carolina State University, Raleigh, NC.
- SNEDECOR, G. W., and W. G. COCHRAN, 1989 *Statistical Methods*. Ames: Iowa State University Press, 8th ed. edition.
- SOBEL, E., and K. LANGE, 1996 Descent graphs in pedigree analysis: applications to haplotyping. *American Journal of Human Genetics* **58**: 1323–1337.
- SPELMAN, R. J., W. COPPIETERS, L. KARIM, J. VAN ARENDONK, and H. BOVENHUIS, 1996 Quantitative trait loci analysis for five milk production traits on chromosome six in the dutch holstein-friesian population. *Genetics* **144**: 1799–1808.

- TAPADAR, P., S. GHOSH, and P. MAJUMDER, 2000 Haplotyping in pedigrees via a genetic algorithm. *Hum. Hered.* **50**: 43–56.
- THOMAS, A., A. GUTIN, V. ABKEVICH, and A. BANSAL, 2000 Multilocus linkage analysis by blocked gibbs sampling. *Statistics and Computing* **10**: 259–269.
- THOMPSON, E., 2007 *Handbook of Statistical Genetics*, chapter Linkage Analysis. Wiley, 3rd edition, 1141–1168.
- THOMPSON, E. A., 2000 *Statistical Inference from Genetic Data on Pedigrees*, volume 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Beachwood, Ohio.
- TZENG, J., and D. ZHANG, 2007 Haplotype-based association analysis via variance component score test. *The American Journal of Human Genetics* **81**: 927–938.
- VANRADEN, P. M., A. H. SANDERS, M. E. TOOKER, R. H. MILLER, H. D. NORMAN, *et al.*, 2004 Development of a national genetic evaluation for cow fertility. *J. Dairy Sci.* **87**: 2285–2292.
- VANRADEN, P. M., and G. R. WIGGANS, 1991 Derivation, calculation, and use of national animal model information. *Journal of Dairy Science* **74**: 2738–2746.
- VISSCHER, P. M., R. THOMPSON, and C. S. HALEY, 1996 Confidence intervals in qtl mapping by bootstrapping. *Genetics* **143**: 1013–1020.
- WAAGEPETERSEN, R., and D. SORENSEN, 2001 A tutorial on reversible jump mcmc with a view toward applications in qtl-mapping. *International Statistical Review / Revue Internationale de Statistique* **69**: 49–61.
- WANG, S., C. J. BASTEN, and Z.-B. ZENG, 2007 Windows qtl cartographer 2.5.
- WANG, T., R. FERNANDO, S. V. D. BEEK, M. GROSSMAN, and J. V. ARENDONK, 1995 Covariance between relatives for a marked quantitative trait locus. *Genetics Selection Evolution* **27**: 251–274.

- WEEKS, D. E., E. SOBEL, J. R. OCONNELL, and K. LANGE, 1995 Computer-programs for multilocus haplotyping of general pedigrees. *Am. J. Hum. Genet.* **56**: 1506–1507.
- WELLER, J. I., Y. KASHI, and M. SOLLER, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J. Dairy Sci.* **73**: 2525–2537.
- WIGGANS, G. R., and P. M. VANRADEN, 1989 Usda-dhia animal model genetic evaluations. Web.
- WIJSMAN, E. M., 1987 A deductive method of haplotype analysis in pedigrees. *American Journal of Human Genetics* **41**: 356–373.
- WINDIG, J. J., and T. H. MEUWISSEN, 2004 Rapid haplotype reconstruction in pedigrees with dense marker maps. *Journal of Animal Breeding and Genetics* **121**: 26–39.
- WRIGHT, S., 1968 *Evolution and Genetics of Populations, Vol. 1, Genetic and Biometric Foundations*, volume 1. University of Chicago Press, Chicago, IL.
- XU, S., 1998 Further investigation on the regression method of mapping quantitative trait loci. *Heredity* **80**: 364–373.
- XU, S., and W. R. ATCHLEY, 1995 A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**: 1189–1197.
- YI, N., B. S. YANDELL, G. A. CHURCHILL, D. B. ALLISON, E. J. EISEN, *et al.*, 2005 Bayesian Model Selection for Genome-Wide Epistatic Quantitative Trait Loci Analysis. *Genetics* **170**: 1333–1344.
- ZENG, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *PNAS USA* **90**: 10972–10976.
- ZENG, Z.-B., 1994 Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

- ZENG, Z.-B., 2000 Statistical methods for mapping quantitative trait loci. Lecture Notes.
- ZHANG, K., F. SUN, and H. ZHAO, 2005 Haplore: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics* **21**: 90–103.
- ZHANG, Q., D. BOICHARD, I. HOESCHELE, C. ERNST, A. EGGEN, *et al.*, 1998 Mapping Quantitative Trait Loci for Milk Production and Health of Dairy Cattle in a Large Outbred Pedigree. *Genetics* **149**: 1959–1973.
- ZOU, F., J. P. FINE, J. HU, and D. Y. LIN, 2004 An Efficient Resampling Method for Assessing Genome-Wide Statistical Significance in Mapping Quantitative Trait Loci. *Genetics* **168**: 2307–2316.
- ZOU, F., J. P. FINE, and B. S. YANDELL, 2002 On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**: 61–75.

APPENDICES

A

Multiple Family Multiple Interval Mapping

In this appendix, we present derivations of the expectation-maximization algorithm to obtain maximum likelihood estimates for the purpose of testing and estimation of QTL effects for a multiple family multiple interval model.

A.1 Model and Notations

In a similar way as the Haley-Knott least squares method being extended to multiple unrelated half-sib families [HOESCHELE *et al.*, 1997; HOESCHELE, 2007], our wMIM can also be extended to analyze multiple families, still assuming additive QTL effects only:

$$y_{ij} = \mu_i + \sum_{r=1}^t b_{ri} x_{ijr}^* + e_{ij} \quad (\text{A.1.1})$$

where

- m is the number of half-sib families (*a.k.a.*, sires);
- n_i is the number of individuals in i -th family;
- t is number of QTL in the model;

- b_{ri} is the QTL additive effects of the r -th QTL;
- x_{ijr}^* denotes the indicator genotype at the r -th locus in the j -th offspring of sire i ;
- e_{ij} is the residual assumed to be normally distributed but with family-specific variances σ_i^2 .

A homoscedastic residual variances across families can also be assumed, which involves the statistical issue of the so-called testing on heteroscedasticity.

A.2 Complete Data Likelihood

Now define the QTL genetic design matrix \mathbf{D} (Cockerham model, [KAO *et al.*, 1999]) as

$$\mathbf{D} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \cdots & d_{1k} & \cdots & \frac{1}{2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{q1} & d_{q2} & \cdots & d_{qk} & \cdots & d_{qm} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ -\frac{1}{2} & -\frac{1}{2} & \cdots & d_{2^t k} & \cdots & -\frac{1}{2} \end{bmatrix}_{2^t \times t}, \quad (\text{A.2.1})$$

and QTL effect matrix \mathbf{E} as

$$\mathbf{E} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1i} & \cdots & b_{1t} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{ki} & \cdots & b_{km} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{t1} & b_{t2} & \cdots & b_{ti} & \cdots & b_{tm} \end{bmatrix}_{t \times m}. \quad (\text{A.2.2})$$

Let \mathbf{D}_q denote the q -th row of \mathbf{D} matrix ($1 \times t$), \mathbf{D}^k is the k -th column of \mathbf{D} ($2^m \times 1$), \mathbf{E}^i is the i -th column of \mathbf{E} matrix ($t \times 1$). Note $d^2 = d_{qk}^2 = 1/4$ for any q and k because only QTL main effects are included in the model.

Further, let p_{ijr} denote the conditional probability at the r -th QTL. Under the assumption that the genotypes of QTL are independent of one another, $p_{ijr} = p(Q_{ri}|\mathbf{M}_{ij}, \theta_r)$ where \mathbf{M}_{ij} is the marker data of offspring j of sire i and θ_r is the position the k -th QTL on the marker map.

The individual likelihood of the j -th individual in family i is:

$$L_{ij}(\mu_i, \mathbf{E}^i, \sigma_i^2 | y_{ij}, \mathbf{M}_{obs}) = \sum_{q=1}^{2^t} p_{ijq} f(y_{ij}; \mu_i + \mathbf{D}_q \mathbf{E}^i, \sigma_i^2). \quad (\text{A.2.3})$$

where $f(y; \mu, \sigma^2)$ is the probability density function of normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The whole log likelihood of all individuals is

$$\ln L(\boldsymbol{\mu}, \mathbf{E}, \boldsymbol{\sigma}^2 | \mathbf{y}, \mathbf{M}_{obs}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \ln \left\{ \sum_{q=1}^{2^t} p_{ijq} f(y_{ij}; \mu_i + \sum_{k=1}^t d_{qk} b_{ki}, \sigma_i^2) \right\}. \quad (\text{A.2.4})$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_m^2)$, $\mathbf{b}_k = (b_{k1}, \dots, b_{km})$.

A.3 E-M Algorithm

A.3.1 The $[s + 1]$ -th E-step:

$$\pi_{ijq}^{[s+1]} = \frac{\pi_{ijq}^{[s]} f(y_{ij}; \mu_i + \mathbf{D}_q \mathbf{E}^i, \sigma_i^2)}{\sum_{q=1}^{2^t} \pi_{ijq}^{[s]} f(y_{ij}; \mu_i + \mathbf{D}_q \mathbf{E}^i, \sigma_i^2)} \quad (\text{A.3.1})$$

where all the parameter estimates are obtained from the $[s]$ -th M-step below.

Define the following vectors and matrices:

$$\mathbf{Y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}_{n_i \times 1}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n_i \times 1}, \quad \mathbf{\Pi}_i = \begin{bmatrix} \pi_{i11} & \pi_{i12} & \cdots & \pi_{i12^m} \\ \pi_{i21} & \pi_{i22} & \cdots & \pi_{i22^m} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{in_i1} & \pi_{in_i2} & \cdots & \pi_{in_i2^m} \end{bmatrix}_{n_i \times 2^m}$$

and that \mathbf{M}' means the transposition of matrix \mathbf{M} .

A.3.2 The $[s]$ -th M-step:

$$\frac{\partial \ln L}{\partial \mu_i} = \sum_{j=1}^{n_i} \sum_{q=1}^{2^t} \pi_{ijq} \frac{y_{ij} - \mu_i - \sum_{k=1}^t d_{qk} b_{ki}}{\sigma_i^2} \equiv 0 \quad (\text{A.3.2})$$

$$\Rightarrow \hat{\mu}_i = \frac{\sum_{j=1}^{n_i} \left[y_{ij} - \sum_{q=1}^{2^t} \pi_{ijq} \left(\sum_{k=1}^t d_{qk} b_{ki} \right) \right]}{n_i} \quad (\text{A.3.3})$$

$$\Rightarrow \hat{\mu}_i = \mathbf{1}'(\mathbf{Y}_i - \mathbf{\Pi}_i \mathbf{D} \mathbf{E}^i) / n_i. \quad (\text{A.3.4})$$

$$(\text{A.3.5})$$

$$\frac{\partial \ln L}{\partial b_{ri}} = \sum_{j=1}^{n_i} \sum_{q=1}^{2^t} \pi_{ijq} \frac{y_{ij} - \mu_i - \sum_{k=1}^t d_{qk} b_{ki}}{\sigma_i^2} d_{qr} \equiv 0 \quad (\text{A.3.6})$$

$$\Rightarrow \hat{b}_{ri} = \frac{\sum_{j=1}^{n_i} \sum_{q=1}^{2^t} \pi_{ijq} \left(y_{ij} - \mu_i - \sum_{k=1, k \neq t}^t d_{qk} b_{ki} \right) d_{qr}}{n_i d^2} \quad (\text{A.3.7})$$

$$\Rightarrow \hat{b}_{ri} = \frac{\mathbf{1}' \left\{ (\mathbf{Y}_i - \mu_i \mathbf{1}) \# (\mathbf{\Pi}_i \mathbf{D}^r) - \mathbf{\Pi}_i \left[\left(\mathbf{D}^{(-r)} \mathbf{E}_{(-r)}^i \right) \# \mathbf{D}^r \right] \right\}}{n_i d^2} \quad (\text{A.3.8})$$

$$\Rightarrow \hat{\mathbf{E}}^i = \text{diag}(\mathbf{V}_i)^{-1} \left[\mathbf{D}' \mathbf{\Pi}_i' (\mathbf{Y}_i - \mu_i \mathbf{1}) - \text{nondiag}(\mathbf{V}_i) \mathbf{E}^i \right] \quad (\text{A.3.9})$$

where $\mathbf{D}^{(-r)}$ means the \mathbf{D} matrix after removal of the r -th column and $\mathbf{E}_{(-r)}^i$ is the i -th column vector of $\mathbf{E}_{(-r)}$ matrix where the r -th row is removed.

\mathbf{V}_i is a matrix of dimension $m \times m$ with the (r, s) element as $\mathbf{1}'\mathbf{\Pi}_i(\mathbf{D}^r \# \mathbf{D}^s)$. $\text{diag}(\mathbf{V}_i)$ is a vector comprised of the diagonal elements of \mathbf{V}_i . $\text{nondiag}(\mathbf{V}_i)$ is a matrix containing all the off-diagonal elements but the diagonal elements are all zero.

$$\frac{\partial \ln L}{\partial (\sigma_i^2)} = \sum_{j=1}^{n_i} \sum_{q=1}^{2^t} \pi_{ijq} \left[\frac{\left(y_{ij} - \mu_i - \sum_{k=1}^t d_{qk} b_{ki} \right)^2}{2 (\sigma_i^2)^2} - \frac{1}{2 (\sigma_i^2)} \right] \equiv 0 \quad (\text{A.3.10})$$

$$\Rightarrow \hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} \sum_{q=1}^{2^t} \pi_{ijq} \left(y_{ij} - \mu_i - \sum_{k=1}^t d_{qk} b_{ki} \right)^2}{n_i} \quad (\text{A.3.11})$$

$$\Rightarrow \hat{\sigma}_i^2 = \frac{(\mathbf{Y}_i - \mu_i \mathbf{1})' (\mathbf{Y}_i - \mu_i \mathbf{1}) - 2 (\mathbf{Y}_i - \mu_i \mathbf{1})' \mathbf{\Pi}_i \mathbf{D} \mathbf{E}^i - \mathbf{1}' [\mathbf{\Pi}_i ((\mathbf{D} \mathbf{E}^i) \# (\mathbf{D} \mathbf{E}^i))]}{n_i} \quad (\text{A.3.12})$$

$$\Rightarrow \hat{\sigma}_i^2 = \frac{1}{n_i} [(\mathbf{Y}_i - \mu_i \mathbf{1})' (\mathbf{Y}_i - \mu_i \mathbf{1}) - 2 (\mathbf{Y}_i - \mu_i \mathbf{1})' \mathbf{\Pi}_i \mathbf{D} \mathbf{E}^i - (\mathbf{E}^i)' \mathbf{V}_i \mathbf{E}^i] \quad (\text{A.3.13})$$

A.3.3 REL is available

In case the reliability (REL) information is available and needs to be incorporated into the model (A.1.1), the E-M formulae will need to be appropriately modified to accommodate the vector of REL, $\boldsymbol{\gamma}$, in the same way as shown in equations (3.3.7).

A.4 Discussions About QTL Testing

It would be of primary interest whether the QTL effect parameter(s) is zero, that is, whether the corresponding QTL exist. It could be a test on a particular QTL in a particular family; or a test on a QTL across all families. Under the assumption that each family has its own residual distribution, the following will be easy to see: for the former testing task, it will give the identical test results as single family analysis within that particular family; for the latter, the test statistic is the sum of all individual family-specific test statistic at that locus.

Of course, the above manipulations would not be valid if the residuals are homogeneous. The question then is, how to justify the assumption of homoscedasticity

or heteroscedasticity in residual variances? Visual examination of the spread of trait data across all families or within a single family might help us better appreciate some aspects of the data, but it would not be conclusive because if there are indeed one or several QTL segregating in the data, then the distribution of the trait data, either across all families or within a single family, would actually be a mixture of several normal components. The spread of the trait data is therefore not only determined by the variation of each component, but also the relative mean value of each component. Also because of the mixture modeling for QTL mapping, there is no validated methods yet for testing between homoscedasticity and heteroscedasticity of multiple family trait phenotype data.

To date, Windows QTL Cartographer does not have appropriate functions for importing and in-memory processing of multiple family data for QTL mapping analysis. It is more of an issue of computer programming, rather than a task that is not feasible to fulfill.

B

Approximation of Mixed-Mixture Model into Mixed Model

B.1 Mixed Effects Mixture Model

Suppose a population consisted of many unrelated nuclear families, each of which has been genotyped and phenotyped according to certain designs. The goal of QTL mapping is to link the variations between trait phenotypes and genotypes and find the most influential locus/loci. Given a QTL at a certain locus, the allele substitution effect in one parent may differ from another, due to either different pairs of allele variants or the uncertain direction of the allele substitution effect (with respect to promoting or demoting trait performances). To better model QTL segregation and effects variation in such populations, we propose to model QTL effects as random effects on the level of parents. QTL substitution effects in all families are assumed to be random samples from a common normal distribution, that is, a particular QTL effect in a family can be regarded as a realized sample from the population distribution of QTL effects, namely $\mathcal{N}(\beta_Q, \sigma_Q^2)$. Within a family, the “realized” random QTL effect(s) are regarded as conditionally fixed effects.

Figure B.1 shows a diagram of our modeling of QTL effects as described above. The new feature in our proposal is, however, the mean of the normal distribution

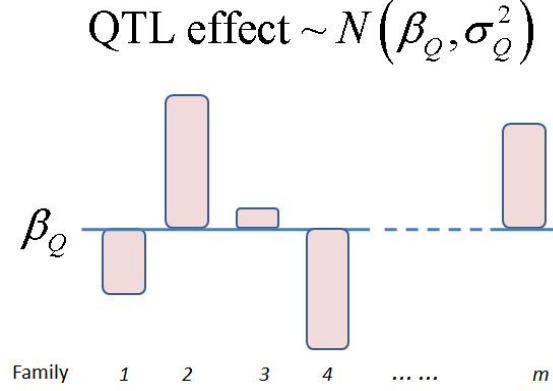


Figure B.1: Diagram of mixed QTL effects in multiple family populations. The QTL effects are random samples from $\mathcal{N}(\beta_Q, \sigma_Q^2)$ but only at the level of families (or parents in families) with the mean, β_Q , not necessarily zero.

(β_Q) is assumed not necessarily to be zero, which reflects the across-family mean of QTL effects in these families, such that the variation of QTL effects observed across families would be better accounted for after this mean effect be addressed.

B.1.1 Model specifications and notations

From the modeling assumptions and specifications above, a mixed effects model of QTL effects can be specified as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^* + b_{i0} + b_{i1} x_{ij}^* + e_{ij} \quad (\text{B.1.1})$$

where

- y_{ij} is the phenotypic value of j -th offspring of the sire i ;
- β_0 and β_1 are the overall intercept and QTL fixed effect;
- x_{ij}^* is the QTL genotype indicator and is defined as $x_{ij}^* = +1/2$ if the QTL allele Q_{i1} is inherited into offspring j , and $-1/2$ if the QTL allele Q_{i2} is inherited into offspring j ;
- b_{i0} and b_{i1} are the random intercepts and random QTL effect in family i , for

which we have $\mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim \mathcal{N}(0, \Delta)$ where $\Delta = \begin{bmatrix} \delta_{00} & \delta_{01} \\ \delta_{01} & \delta_{11} \end{bmatrix}$ with δ_{11} being the variance of QTL random effect;

- e_{ij} is the residual assumed to be *i.i.d.* distributed as $\mathcal{N}(0, \sigma_e^2)$.

For consistency in notations with those in Chapter 4, we have replaced β_Q and σ_Q^2 with β_1 and δ_{11} , respectively.

The model formula (B.1.1) for family i can then be written as follows:

$$\mathbf{y}_i = H_i^* \boldsymbol{\beta} + H_i^* \mathbf{b}_i + \mathbf{e}_i \quad (\text{B.1.2})$$

where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}_{n_i \times 1}, \quad H_i^* = \begin{bmatrix} 1 & x_{i1}^* \\ 1 & x_{i2}^* \\ \vdots & \vdots \\ 1 & x_{in_i}^* \end{bmatrix}_{n_i \times 2}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix}, \quad \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}_{n_i \times 1}.$$

Finally, in matrix notation, the model of the whole data is given below:

$$\mathbf{Y} = \mathbf{X}^* \boldsymbol{\beta} + \mathbf{Z}^* \mathbf{b} + \mathbf{e} \quad (\text{B.1.3})$$

with \mathbf{Y} and \mathbf{e} being vectors of trait and residuals and

$$\mathbf{X}^* = \begin{bmatrix} H_1^* \\ H_2^* \\ \vdots \\ H_m^* \end{bmatrix}_{N \times 2}, \quad \mathbf{Z}^* = \begin{bmatrix} H_1^* & & & \\ & H_2^* & & \\ & & \ddots & \\ & & & H_m^* \end{bmatrix}_{N \times 2m}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}_{2m \times 1}$$

where $N = \sum_{i=1}^m n_i$, the total number of observations, and

$$\begin{aligned}\mathbf{b} &\sim \mathcal{N}(0, \Delta \otimes \mathbf{I}_m) \equiv \mathcal{N}(0, \underbrace{\text{diag}\{\Delta, \dots, \Delta\}}_m), \\ \mathbf{e} &\sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}).\end{aligned}$$

Note the dimension of the identity matrix \mathbf{I} for the residuals is $N \times N$. However, this subscript of dimension for identity matrices will be omitted whenever it can be clearly determined from the context.

B.1.2 Mixture of variance structures

The variance-covariance structure of the model (B.1.3) is block diagonal with each block the variance-covariance matrix of the respective family, which is a corollary of the assumption of unrelated families. For the i -th family, its variance-covariance matrix is given below:

$$\begin{aligned}\mathbf{V}_i^* &= \mathbf{H}_i^* \Delta (\mathbf{H}_i^*)' + \sigma_e^2 \mathbf{I}_{n_i \times n_i} \\ &= \begin{bmatrix} 1 & x_{i1}^* \\ 1 & x_{i2}^* \\ \vdots & \vdots \\ 1 & x_{in_i}^* \end{bmatrix}_{n_i \times 2} \begin{bmatrix} \delta_{00} & \delta_{01} \\ \delta_{01} & \delta_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{i1}^* & x_{i2}^* & \cdots & x_{in_i}^* \end{bmatrix}_{2 \times n_i} + \sigma_e^2 \mathbf{I} \\ &= \left\{ \left\{ \delta_{00} + \delta_{10} x_{ij}^* + \delta_{01} x_{ij'}^* + \delta_{11} x_{ij}^* x_{ij'}^* \right\}_{[j,j']} \right\}_{j=1, \dots, n_i; j'=1, \dots, n_i} + \sigma_e^2 \mathbf{I} \quad (\text{B.1.4})\end{aligned}$$

Since the QTL allele of sire origin in the progeny is unobserved, x_{ij}^* is missing data. For n_i individuals, there are in total 2^{n_i} possible different configurations of the joint QTL genotypes $\mathbf{x}_i^* = \begin{bmatrix} x_{i1}^* & \cdots & x_{in_i}^* \end{bmatrix}'$. For simplicity in referring to the

realization of \mathbf{x}_i^* , we define a genetic design matrix for family i , \mathbf{D}_i , as follows:

$$\mathbf{D}_i = \begin{bmatrix} \frac{1}{2} & \cdots & \frac{1}{2} & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{k1} & \cdots & d_{kj} & \cdots & d_{kn_i} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} & \cdots & -\frac{1}{2} \end{bmatrix}_{2^{n_i} \times n_i} . \quad (\text{B.1.5})$$

Further denote D_{ik} the k -th row vector in \mathbf{D}_i that corresponds to the k -th possible realization of QTL genotypes in n_i offspring of family i . Then, under the assumption of independent gametogenesis, we have $P(\mathbf{x}_i^* = D_{ik}) = \prod_{j=1}^{n_i} P(x_{ij}^* = D_{ik,j})$ in which $D_{ik,j}$ denotes the j -th element in D_{ik} .

B.1.3 Complete data likelihood

The complete data likelihood of m unrelated half-sib families, treating \mathbf{x}^* as missing data, takes the following form:

$$\log L(\boldsymbol{\beta}, \Delta, \sigma_e^2 | \mathbf{Y}, \mathbf{M}_{obs}) = \sum_{i=1}^m \sum_{k=1}^{2^{n_i}} \log f(\mathbf{y}_i | \boldsymbol{\beta}, \Delta, \sigma_e^2, \mathbf{x}_i^*) P(\mathbf{x}_i^* = D_{ik}) \quad (\text{B.1.6})$$

This complete data likelihood is the sum of each family-specific likelihood, which itself is a mixture of a large number of components, as is 2^{n_i} , with mixing proportions as $P(\mathbf{x}_i^* = D_{ik})$. The likelihood of each component is derived from a linear mixed model consisting more than one variance components. Although methods are available for maximizing any individual likelihood derived from a linear mixed effects model, there is no theory nor software program for maximizing a weighted sum of such likelihoods. It may occur to one's mind that E-M algorithm may be of use for maximizing this complete data likelihood. However, besides the difficulty due to lack of theory and software tools for maximize a mixture of mixed effects likelihoods, the massively large number of mixture components itself will demand massive amount of computing resources such as computation time and memory and storage capaci-

ties, making it computationally prohibitively expensive. The fact that the number of half-sib offspring of a dairy bull is usually large, in dozens and hundreds, due to the extensive use of artificial insemination as well as the breeding practice makes it imperative to find an approximation to assess the likelihood in equation (B.1.6).

B.2 Approximated Mixed Effects Model

As seen in the previous section, the difficulty in evaluating the complete data likelihood lies in the massive number of mixture components and the challenge that there is no effective way to pinpoint and discard QTL genotype configurations (D_{ik} , $k = 1, \dots, 2^{n_i}$) with small probabilities. Therefore, one strategy of approximation is to reduce or avoid evaluating the mixture model by replacing the unobserved QTL genotypes in the offspring by their conditional expectations given the observed marker data, *i.e.*,

$$\begin{aligned} z_{ij} = E(x_{ij}^* | \mathbf{M}_{obs}) &= \frac{1}{2} \Pr \left(x_{ij}^* = \frac{1}{2} | \mathbf{M}_{obs} \right) + \left(-\frac{1}{2} \right) \Pr \left(x_{ij}^* = -\frac{1}{2} | \mathbf{M}_{obs} \right) \\ &= \Pr \left(x_{ij}^* = \frac{1}{2} | \mathbf{M}_{obs} \right) - \frac{1}{2}. \end{aligned} \quad (\text{B.2.1})$$

The model in (B.1.3) then turns into a linear mixed model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \quad (\text{B.2.2})$$

where the definitions of terms and parameters are the same as before except that the design matrices \mathbf{X} and \mathbf{Z} are the element-wise expectation of \mathbf{X}^* and \mathbf{Z}^* , respectively.

Now the problem has reduced to a linear mixed model at a given locus. Likelihood ratio test (LRT) can be constructed for testing QTL ($H_{02} : \delta_{11} = 0$ vs. $\delta_{11} > 0$). However, the restricted maximum likelihood (REML) method would not be appropriate in such cases because the fixed effects would also be subject to testing. Therefore, the maximum likelihood (ML) method is used throughout the analysis.

The variance of \mathbf{Y} takes the following form:

$$\begin{aligned}
 \text{Var}(\mathbf{Y}) = \mathbf{V} &= \mathbf{Z} \Delta \mathbf{Z}' + \sigma_e^2 \mathbf{I} \\
 &= \begin{bmatrix} \mathbf{V}_1 & & & \\ & \mathbf{V}_2 & & \\ & & \ddots & \\ & & & \mathbf{V}_m \end{bmatrix} \\
 &= \begin{bmatrix} \mathbf{Z}_1 \Delta \mathbf{Z}'_1 + \sigma_e^2 \mathbf{I} & & & \\ & \mathbf{Z}_2 \Delta \mathbf{Z}'_2 + \sigma_e^2 \mathbf{I} & & \\ & & \ddots & \\ & & & \mathbf{Z}_m \Delta \mathbf{Z}'_m + \sigma_e^2 \mathbf{I} \end{bmatrix} \quad (\text{B.2.3})
 \end{aligned}$$

with

$$\begin{aligned}
 \mathbf{V}_i &= \mathbf{Z}_i \Delta \mathbf{Z}'_i + \sigma_e^2 \mathbf{I} \\
 &= \left\{ \left\{ \delta_{00} + \delta_{10} z_{ij} + \delta_{01} z_{ij'} + \delta_{11} z_{ij} z_{ij'} \right\}_{[j,j']} \right\}_{j=1, \dots, n_i; j'=1, \dots, n_i} + \sigma_e^2 \mathbf{I} \\
 &= \delta_{11} \mathbf{z}_i \mathbf{z}'_i + \delta_{01} (\mathbf{1} \mathbf{z}'_i + \mathbf{z}_i \mathbf{1}') + \delta_{00} \mathbf{1} \mathbf{1}' + \sigma_e^2 \mathbf{I}_i. \quad (\text{B.2.4})
 \end{aligned}$$

Note that we have assumed an unstructured variance-covariance structure between the random intercept and QTL effects. Simple algebra leads to the following if independence among random effects (such that $\delta_{01} = \delta_{10} = 0$) is assumed:

$$\begin{aligned}
 \mathbf{V}_i &= \left\{ \left\{ \delta_{00} + \delta_{11} z_{ij} z_{ij'} \right\}_{[j,j']} \right\}_{j=1, \dots, n_i; j'=1, \dots, n_i} + \sigma_e^2 \mathbf{I} \\
 &= \delta_{11} \mathbf{z}_i \mathbf{z}'_i + \delta_{00} \mathbf{1} \mathbf{1}' + \sigma_e^2 \mathbf{I}. \quad (\text{B.2.5})
 \end{aligned}$$