MULTIPLE-TRAIT MULTIPLE-INTERVAL MAPPING OF QUANTITATIVE-TRAIT LOCI

by

ROBY JOEHANES

B.S., Universitas Pelita Harapan, Indonesia, 1999
M.S., Kansas State University, 2002

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2009

Approved by:

Major Professor
Dr. Gary L. Gadbury

Copyright

Abstract

QTL (quantitative-trait locus) analysis aims to locate and estimate the effects of genes that are responsible for quantitative traits, such as grain protein content and yield, by means of statistical methods that evaluate the association of genetic variation with trait (phenotypic) variation. Quantitative traits are typically polygenic, *i.e.*, controlled by multiple genes, with varying degrees of influence on the phenotype. Several methods have been developed to increase the accuracy of QTL location and effect estimates. One of them, multiple interval mapping (MIM) (Kao et al. 1999), has been shown to be more accurate than conventional methods such as composite interval mapping (CIM) (Zeng 1994). Other QTL analysis methods have been developed to perform additional analyses that might be useful for breeders, such as of pleiotropy and QTL-by-environment (Q×E) interaction. It has been shown (Jiang and Zeng 1995) that these analyses can be carried out with a multivariate extension of CIM (MT-CIM) that exploits the correlation structure in a set of traits. In doing so, this method also improves the accuracy of QTL location detection. This thesis describes the multivariate extension of MIM (MT-MIM) using ideas from MT-CIM. The development of additional multivariate tests, such as of pleiotropy and Q×E interaction, and several methods pertinent to the development of MT-MIM are also described. A small simulation study shows that MT-MIM is more accurate than MT-CIM and univariate MIM. Results for real data show that MT-MIM is able to provide a more accurate and precise estimate of QTL location.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my major professor, Dr. Gary Gadbury, for his guidance and patience. Without his guidance and questions, this thesis would not have been possible.

I would like to thank my committee member and my major professor in Genetics, Dr. James C. Nelson, for his guidance and correction in my thesis. He has given much influence in my graduate education.

I would like to thank Dr. Suzanne Dubnicka, another member of my committee, for her technical insights and teachings of basic statistical knowledge.

I would like to thank Dr. Paul Nelson for his helpful comments on my draft.

Dedication

I dedicate this report to my wife, Elkarisma, who has made countless sacrifices to make this report possible.

Chapter 1

Introduction to quantitative trait locus (QTL) analysis

QTL analysis aims to locate and estimate the effects of genes that are responsible for quantitative traits, such as grain protein content and yield, by means of statistical methods that evaluate the association of genetic variation with trait (phenotypic) variation. Quantitative traits are typically polygenic, *i.e.*, controlled by multiple genes, with varying degrees of influence on the phenotype.

Several methods have been developed to increase the accuracy of QTL location and effect estimates. One of them, multiple interval mapping (MIM) (Kao et al. 1999), has been shown to be more accurate than conventional methods such as composite interval mapping (CIM) (Zeng 1994).

Other QTL analysis methods have been developed to perform additional analyses that might be useful for breeders, such as of pleiotropy and QTL-by-environment (Q×E) interaction. It has been shown (Jiang and Zeng 1995) that these analyses can be carried out with a multivariate extension of CIM (MT-CIM) that exploits the correlation structure in a set of traits. In doing so, this method also improves the accuracy of QTL location detection.

This thesis describes the multivariate extension of MIM (MT-MIM) using ideas from MT-CIM. The development of additional multivariate tests, such as of pleiotropy and Q×E interaction, and several methods pertinent to the development of MT-MIM are also de-

scribed.

In order to understand QTL mapping, it is necessary to know how genetic variation arises. This is explained in the following sections.

1.1 Chromosome, locus, allele, genotype

A chromosome can be thought of as an array of ordered *loci* (singular *locus*). Each locus can be thought of as a variable that can take any of several discrete values called *alleles*. For species we consider here, chromosomes come in pairs for which the same loci lie in the same order. For example, humans have 23 pairs, yeast 16, and barley 7. Here we consider one such pair.

The two alleles at a locus on a chromosome pair are called the locus *genotype*. If a genotype has identical alleles, it is *homozygous*. Otherwise, it is *heterozygous*.

For example, consider a chromosome pair with only one locus, A. Suppose for this locus there are only two alleles, A and a, in a population of chromosomes. In this case, there are three possible genotypes: AA, Aa, and aa. AA and aa are homozygous, while Aa is a heterozygous genotype. Genotype aA is indistinguishable from genotype Aa.

1.2 Gametogenesis and meiosis

Genotypic variation is created from a process called *crossing over*. In crossing over, the paired chromosomes exchange segments, as shown in Figure 1.1, and produce two child chromosomes, or *gametes*. The result of this process is called *recombination*. The event in which crossing over occurs is called *meiosis*. Although meiosis takes place in all plant and animal sex organs, recombination can be detected only when the chromosome segments exchanged carry different alleles.

Gametogenesis occurs in both father and mother. One gamete from the father will pair with one from the mother to form the progeny chromosome pair.

1.3 Recombination fraction

There is a certain probability of crossing over between every two loci. Consider a pair of chromosomes with two loci, A and B. This pair forms a new gamete as shown in Figure 1.2. The parental chromosomes have genotypes $AaBb$. The genotype of the gamete is one of AB, ab, Ab, or aB. AB and ab are called *parental* types because the loci are the same as those of the parents, while Ab and aB are *recombinant* types. Recombinant types are the result of crossing over.

Although the true probability of crossing over is not known, it can be estimated from the proportion of recombinant types observed in a mating experiment. This proportion is called the *recombination fraction*, or r. Let f_{XY} be the frequency of gamete XY. The recombination fraction between locus A and B, or r_{AB} in this example is then $r_{AB} = (f_{Ab} + f_{aB})/(f_{AB} + f_{ab} + f_{Ab} + f_{aB})$.

The smaller the recombination fraction, the lower is the probability of a crossover. When r is very small, the two loci are said to be tightly *linked*. When the two loci are unlinked, the genotypes of these loci are independently sampled during gametogenesis. In this case, asymptotically, the frequency of parental types and of recombinant types are the same, or $r = 0.5$.

For visualization of markers in a genetic map, it is convenient to cast recombination

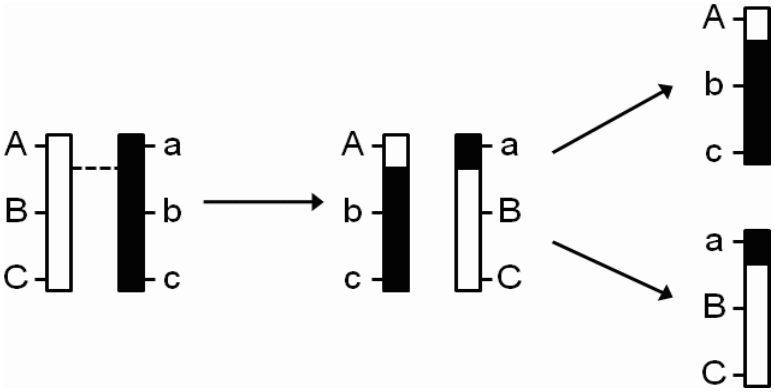


Figure 1.1: *A pair of chromosomes exchange segments via crossing over. The broken line shows the crossover point.*

fractions as *genetic distance*s. Since these fractions do not have an additive relationship, *i.e.*, for loci A, B, and C, $r_{AB} + r_{BC} \neq r_{AC}$, a *mapping function* is required to convert them into distances to preserve additivity. With one of these functions, recombination fractions between pairs of loci can be used to construct a *genetic map*. In genetic maps, genetic distances are expressed in terms of *Morgan*s (M) or *centiMorgan*s (cM), after geneticist Thomas Hunt Morgan.

1.4 Mapping population and mating design

A mapping population for QTL analysis starts with a cross of two parents that have contrasting values for traits of interest. These parents ideally have homozygous genotypes at all loci. In plants such parents are created by repeated self-pollination, or *selfing*, over multiple generations. These parents are known as *inbred lines*, or simply *lines*.

A *mating design* is a description of the crosses used to produce recombinant progeny starting with the inbred parents. Its goal is to create genotypic variation (arising from random parental chromosome sampling) at each locus, and recombination (arising from crossing over) between loci. The first governs the accuracy of QTL effect estimates and the second that of QTL location estimates.

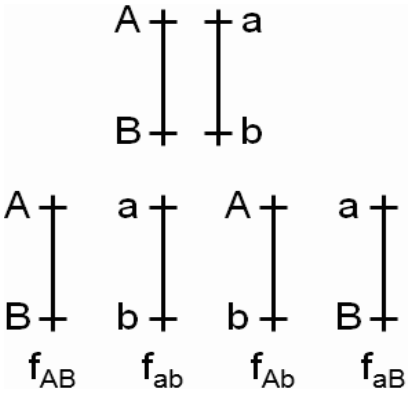


Figure 1.2: *The assortment of gametes produced in gametogenesis. The upper half shows the parental chromosome pair, with genotype AaBb; the lower half shows the array of gametes that it can produce. AB and ab represent parental and Ab and aB recombinant gametes. $f(XY)$ denotes the frequency of gamete XY.*

First, the inbred parents are crossed to form an F_1 progeny. When both parents are all homozygous with differing alleles at every locus, the F_1 progeny are all heterozygous. Then, crosses may be carried out either to self each F_1 plant to form F_2 progeny (Figure 1.3(a)), or to cross the F_1 with one of the parents to form *backcross* (BC_1) progeny. In the F_2, there are two recombinationally informative meioses per progeny, one for each parent, while in the BC_1 there is only one, in the F_1 parent.

Recombinant inbred lines (RILs) are another commonly used mating design for QTL analysis, in which repeated selfings are performed over several (usually 6–8) generations from the F_1. The repeated selfings yield progeny with homozygous genotypes at almost every locus. There are two recombinationally informative meioses per progeny per generation, one in each parent per progeny per generation.

Haploid doubling (DH) is also a common mating design. DH creates an instant homozygote through methods such as anther culture, in which the male gametes of an F_1 progeny are cultured and then treated with a chemical to induce doubling. In DH, there is only one meiosis per progeny, *i.e.*, from the F_1 parent. A DH progeny, like a RIL, can be selfed to create genetically identical offspring for replicated experiments.

Each mating design produces different proportions of genotypes for each locus. These proportions are the unconditional genotypic probability of each locus. For example, in an

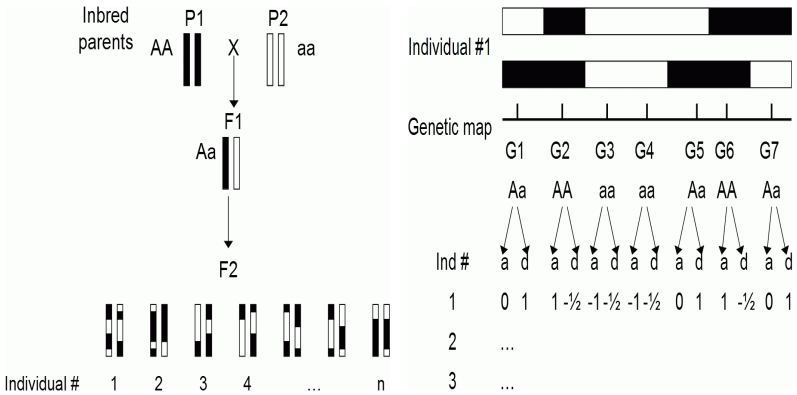


Figure 1.3: *(a) F_2 mating design; (b) Converting genotype data to numerical values*

F_2 population, the expected frequency distribution of aa, Aa, and AA genotypes is 1:2:1. In other words, the genotypic probability distribution for all loci is 0.25, 0.5, and 0.25, for aa, Aa, and AA, respectively. In the BC_1, the ratio is 1:1:0, and in the RIL and DH, it is 1:0:1.

Table 1.1: *Conversion table for common mating designs. The numbers in column A describe a contrast, i.e., they sum to 0. The numbers in column D describe a contrast corrected for the Aa genotype because Aa and aA are indistinguishable.*

Mating Design	**Column A**			**Column D**		
	AA	***Aa***	***aa***	***AA***	***Aa***	***aa***
F_2	1	0	-1	-0.5	0.5	-0.5
BC_1	0	0.5	-0.5	N/A		
RIL / DH	1	0	-1	N/A		

1.5 Genotypic assay and conversion to numerical values

After the mapping population is formed, the genotypes of the progeny at preselected loci are assayed. The loci upon which genotypes are assayed are called DNA *markers*.

After the genotypes from all progeny are assayed, they are converted to numeric form according to Table 1.1 and Figure 1.3(b). The conversion is done by multiplication of the genotype data by a suitable contrast. For example, for an F_2 population, the contrast in column A is used to measure an *additive* effect, which is the difference between the phenotypic means of the two homozygous genotypes AA and aa. The contrast in column D is used to measure a *dominance* effect, which is the difference between the phenotypic means of the Aa genotype and the average of the homozygotes. When only two genotypes are present, such as only Aa and aa in a BC_1 design, only the additive effect can be estimated. These values are used in QTL mapping as the values of the explanatory variables.

1.6 Summary of a QTL mapping process

In summary, QTL mapping is done as follows:

- Select two inbred lines that have contrasting values in traits of interest.
- Make a mapping population from these lines with a suitable mating design.
- If a genetic map is available, select a set of loci as markers that are dense enough to cover the entire map. If a genetic map is not available, there are methods to create and select markers from which a genetic map can be constructed.
- Assay the genotypes of each progeny at the selected markers.
- Assay the phenotypes or traits of interest.
- Use an algorithm to search for QTLs.

Chapter 2

Review of existing QTL mapping methods

2.1 Method overview

The statistical model used for QTL analysis is usually a general linear model (GLM), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ represents quantitative trait data (*e.g.*, plant height), $\mathbf{X}$ represents the genotypic data as described previously, $\boldsymbol{\beta}$ represents QTL effects, and $\boldsymbol{\epsilon}$ represents residuals. These are usually assumed to be independent, with $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix, with n the number of observations. Here, $\mathbf{y}$ is an $n \times 1$ vector, $\mathbf{X}$ is an $n \times (cg+1)$ matrix, and $\boldsymbol{\beta}$ is a $(cg+1) \times 1$ vector, where c is the number of effects per marker and g is the number of markers in the model. In the F_2 design, $c = 2$ (*e.g.*, additive and dominance effects), while in BC_1 and RIL, $c = 1$. The number g of markers in the model varies according to the QTL-mapping method. In addition to the genotypic values, the $\mathbf{X}$ matrix may contain non-genetic factors whose fixed effects are of interest. The structure of $\mathbf{X}$ varies with the method.

In its simplest form (*i.e.*, single-marker regression, or SMR), the QTL analysis uses the genotypic data of one marker at a time ($g = 1$). At each marker, a statistic, typically a LOD score, is computed. The LOD score is a log to base 10 of a likelihood ratio test (LRT). In this test, the null hypothesis is that the marker has neither additive nor dominance effect, while the alternative hypothesis is the negation of the null hypothesis. This process is repeated

for all m markers; that is, the model is fitted to each marker, resulting in m tests. If an association between an existing marker and the trait is detected, *i.e.*, if the marker has a LOD score exceeding some predefined threshold, it is inferred that a QTL is located near that marker based on the observation that the genotypes of loci that are closer together are more likely to be inherited together. So to increase the odds of detection of a true QTL, and to refine the precision of location estimates, the sampled markers must be adequately densely distributed.

Multiple-testing issues and correlations among the test statistics due to linkage make it difficult to determine a statistically significant threshold for QTLs. The conventional threshold of LOD 3 based on the asymptotic distribution may not necessarily reflect the true significance threshold (Mangin et al. 1994).

The multiple-testing problem can be addressed by false discovery rate (FDR) methods (Benjamini and Hochberg 1995; Storey and Tibshirani 2003) and permutation analyses (Churchill and Doerge 1994). In FDR methods, the rate of false discovery or false positives is estimated for determining the correct cutoff (Benjamini and Yekutieli 2005). In permutation analysis, the LOD score threshold is empirically sampled under the null hypothesis. In general, permutation analysis is the preferred method because of the correlation among the statistics due to linkage.

Inferring the QTL genotype *within* marker intervals may improve QTL detection accuracy, especially when obtaining a dense map of a particular species is not possible for technical and/or economic reasons. It requires a controlled mating so that the prior probability of each genotype for each marker is known. This probability enables the inference of QTL genotypic probability at any given point within marker intervals. Such inference is useful in QTL interval mapping (IM). In QTL IM, each chromosome is divided into equal-sized intervals, measured in units of genetic distance.

Inference of QTL genotype (Lander and Botstein 1989) relies on the posterior probability distribution of QTL genotype given the genotypes of markers flanking the QTL, the

recombination fraction between the QTL and the flanking markers, and the unconditional genotypic probability associated with the mating design. This approach is also useful for computing the genotypic probability distribution of missing marker data.

Consider a QTL Q between markers A and B, with two alleles each in the population. Let A and B denote alleles contributed by the first parent and a and b those by the second. Let d_{XY} and r_{XY} denote the distance and the recombination fraction between X and Y. A mapping function (Haldane 1919) is used to convert the distance into the recombination fraction by $r_{XY} = (1 - \exp(-2d_{XY}))/2$, where d_{XY} is expressed in Morgan. The expected frequency of any gamete can be expressed in terms of recombination fractions. For example, at meiosis in the F_1, the probabilities of the two nonrecombinant gametes AQB and aqb are both $(1 - r_{AQ})(1 - r_{QB})/2$ reflecting the absence of recombination between both AQ and QB. Thus, given the flanking marker genotypes AA and BB, the probability that G_Q, the genotype at QTL Q, equals QQ is

$$P(G_Q = QQ|A = AA, B = BB) = (1 - r_{AQ})^2(1 - r_{QB})^2/4$$

because it is formed by pairing two AQB gametes. Likewise, the probability of Qq and qq genotypes are

$$P(G_Q = Qq|A = AA, B = BB) = (1 - r_{AQ})(1 - r_{QB})r_{AQ}r_{QB}/2$$

and

$$P(G_Q = qq|A = AA, B = BB) = r_{AQ}^2 r_{QB}^2/4$$

The preceding method is practical only for simple mating designs such as BC_1 and F_2. A matrix method accomodates more complex mating designs and incompletely informative marker genotypes. A Markov-chain approach (Jiang and Zeng 1997) was developed as a generalization allowing use of the information from all markers in the chromosome. The computation is as follows. Let $\mathbf{p}_k^L$ and $\mathbf{p}_k^R$ be the probability of QTL k, given the markers to its left and to its right. Let $A\#B$ denote the Hadamard (componentwise) product of

two vectors, A and B. Let $\mathbf{q}_k$ be the unconditional genotypic probability from the mating population. The expected frequency of each genotype of QTL k is a 3×1 vector given by $\mathbf{v} = \frac{\mathbf{q}_k \# (\mathbf{p}_k^L \# \mathbf{p}_k^R)}{\mathbf{q}_k' (\mathbf{p}_k^L \# \mathbf{p}_k^R)}$. This probability vector is then multiplied by the contrast vectors in Table 1.1, as illustrated in Figure 2.1.

The inference of QTL genotypic probability within marker intervals is required for QTL interval mapping (IM). In effect, IM interpolates the LOD score within marker intervals. The points within the intervals are treated as "virtual markers" with all-missing data. The Markov-chain approach is used to infer the QTL probability distribution, which is a 3×1 vector denoting the respective probabilities of AA, Aa, and aa genotypes.

Simple interval mapping (SIM) is the IM analog of SMR. Just as in SMR, a statistic (LOD) is computed at each test position, instead of at each marker. SIM may detect a few more QTLs than SMR, as shown in Figure 2.2.

Composite interval mapping (CIM) (Zeng 1994; Jansen 1994; Jansen and Stam 1994) improves upon SIM by including in the model *background* markers or *cofactor*s, *i.e.*, markers in the genetic map that are selected to reduce residual genetic variation arising from QTLs not linked to the QTL being tested. Such reduction allows more precise QTL location estimation by giving narrower peaks. Multiple interval mapping (MIM)(Kao et al. 1999) improves upon CIM by fitting multiple putative QTLs simultaneously via an EM algorithm.

Bayesian interval mapping (BIM) (Satagopan et al. 1996) uses a Markov-chain Monte Carlo (MCMC) approach to sample the posterior probabilities of the QTL genotypes, effects, and locations. These are accepted with a proposal probability computed with a Metropolis-Hastings (MH) method. An improvement of BIM using reversible-jump MCMC (RJMCMC) was proposed (Sillanpää and Arjas 1998; Stephens and Fisch 1998) to address the model-selection problem.

Only SMR, SIM, CIM, and MIM are discussed in the following subsections as pertinent to the development of multiple-trait MIM (MT-MIM) described in this report. BIM and other methods are not discussed. All derivations of formulas are available in the respective

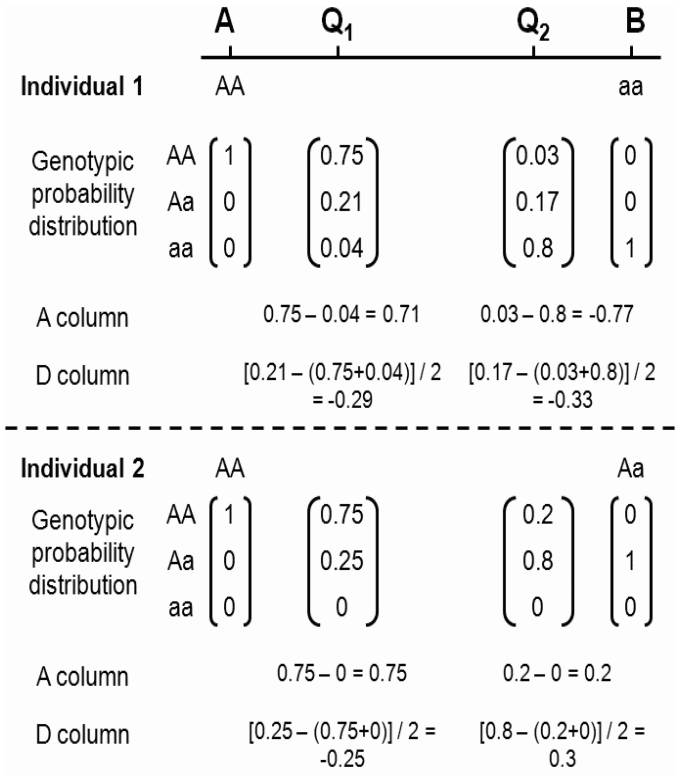


Figure 2.1: *An illustration of a Markov-chain method for inferring QTL genotype. In this example, QTL Q_1 and Q_2 are flanked by markers A and B. The genotypic probability distribution (GPD) vectors of Q_1 and Q_2 are computed with a Markov-chain method. Intuitively, since Q_1 is closer to marker A, its GPD should resemble marker A's. Similarly, since Q_2 is closer to marker B, its GPD should resemble marker B's. The value for the corresponding A column for the QTL is its GPD multiplied by $(-1, 0, 1)'$. Likewise, the value for the corresponding D column for the QTL is its GPD multiplied by $(-0.5, 0.5, -0.5)'$. In all interval-mapping methods, these values replace the values from the substitution rule described in the text. Notice that for individual 2, the values in A and D columns for Q_1 and Q_2 are different because of differing flanking marker genotypes.*

papers and are included for the purpose of MT-MIM development. MT-MIM is discussed in a separate chapter.

2.2 Least-squares-based SMR, SIM, and CIM

In single-marker regression (SMR), the number of markers in the model, g, is one and the model reduces to $\mathbf{y} = \mu\mathbf{1} + a\mathbf{x_a} + d\mathbf{x_d} + \boldsymbol{\epsilon}$. The $\mathbf{1}$ vector signifies a column of ones and μ is an overall mean effect. The $n \times 1$ vectors $\mathbf{x_a}$ and $\mathbf{x_d}$ are genotypic values converted as shown in Table 1.1 and Figure 1.3 (b). The scalars a and d are the additive and dominance effects of interest to geneticists. If the mating design does not have column D in Table 1.1, such as in BC_1 and RIL, the term $d\mathbf{x_d}$ is omitted, since the dominance effect cannot be estimated for that design. The statistics are computed at each marker in turn across the genetic map. An example of an SMR plot appears in Figure 2.2.

Simple interval mapping (SIM) uses the same model used in SMR except that the values

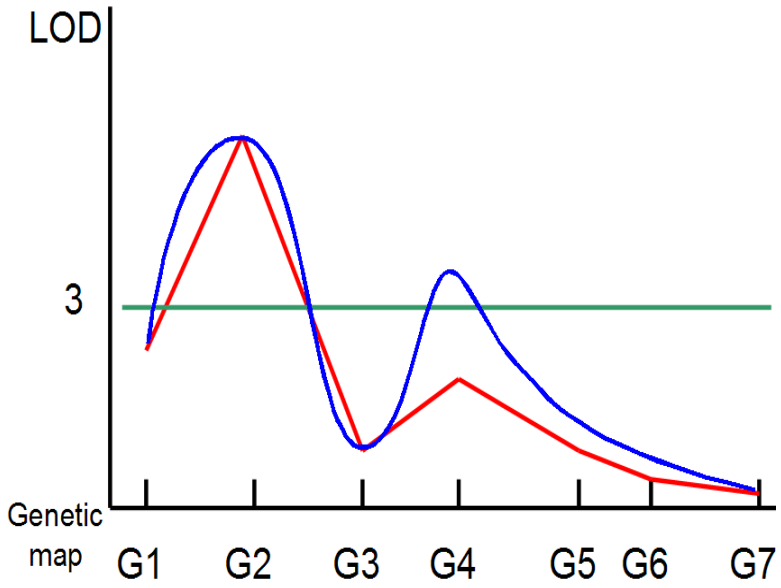


Figure 2.2: *Typical QTL-mapping profile. The single-marker regression (SMR) plot is in red. The simple interval mapping (SIM) plot is in blue. The LOD 3 line represents a significance threshold for a QTL. In this example, both SMR and SIM detect a QTL near marker G_2, but only SIM can detect a QTL near marker G_4.*

in vectors $\mathbf{x_a}$ and $\mathbf{x_d}$ are computed from QTL genotype probability estimates described previously.

In composite interval mapping (CIM), the linear model is $\mathbf{y} = \mu\mathbf{1} + a\mathbf{x_a} + d\mathbf{x_d} + \mathbf{X}^*\beta^* + \boldsymbol{\epsilon}$. The vectors $\mathbf{x_a}$ and $\mathbf{x_d}$ are as in SIM. $\mathbf{X}^*$ is an $n \times cm$ matrix of background markers, where c is the number of QTL effects (*i.e.*, the number of columns in Table 1.1 for the given mating design) and m is the number of background markers. These can be selected manually or through model-selection methods, such as stepwise- or forward-selection. The entries of the $\mathbf{X}^*$ matrix are obtained by the same conversion rule used for converting marker genotypes to numbers in SMR. Thus, in the absence of background marker matrix $\mathbf{X}^*$, CIM reduces to SIM.

SMR, SIM, and CIM models are simple- or multiple-regression models, which can be solved by least-squares methods. Let $\mathbf{X} = [\mathbf{1}|\mathbf{x_a}|\mathbf{x_d}|\mathbf{X}^*]$ and $\boldsymbol{\beta}' = [\mu|a|d|\boldsymbol{\beta}^{*'}]$. Thus, their models can be summarized into $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The solution of the QTL effects is expressed by $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Naïve implementation of a least-squares method is slow and is numerically unstable owing to the finite precision of real numbers in computers. An ordinary least-squares solution involves three matrix multiplications and one matrix inversion. Precision loss occurs at every numerical operation, especially in matrix inversions. Moreover, matrix multiplications and inversions are among the most computationally expensive basic matrix operations.

To improve numerical accuracy and computation speed, QR decomposition is usually used by common statistical software. $\mathbf{X}$ is decomposed into $\mathbf{Q}$ and $\mathbf{R}$, where $\mathbf{Q}$ is an orthogonal matrix (*i.e.*, $\mathbf{Q}' = \mathbf{Q}^{-1}$) and $\mathbf{R}$ is an upper triangular square matrix. Thus:

$$\begin{aligned}\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{R}'\mathbf{Q}'\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} \\ &= (\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} = \mathbf{R}^{-1}\mathbf{R}^{'-1}\mathbf{R}'\mathbf{Q}'\mathbf{y} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y} \\ \mathbf{R}\boldsymbol{\beta} &= \mathbf{Q}'\mathbf{y}\end{aligned}$$

The solution for $\boldsymbol{\beta}$ can be obtained by backward substitution. Numerical accuracy is improved because matrix inversion is no longer necessary and only two multiplications are

needed. Computation speed is improved because partial QR decomposition can be performed for each marker or interval. Result computation from partial decomposition is much faster than from full decomposition.

The null hypothesis is that the QTL effects are zero ($a = d = 0$), *i.e.*, there is no QTL. In SMR and SIM, the null model is $\mathbf{y} = \mu\mathbf{1} + \boldsymbol{\epsilon}$. In CIM, the null model is $\mathbf{y} = \mu\mathbf{1} + \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$. The corresponding F statistics are computed. LOD score can be obtained from LOD $= \frac{n}{2\log 10} \log\left(F \frac{\mathrm{df}_R}{\mathrm{df}_E} + 1\right)$ (Doerge 1995), where n is the number of progeny, df_R and df_E are the degrees of freedom for regression and error, and F is the F statistic.

2.3 EM-based SMR, SIM, and CIM

Single-marker regression (SMR), simple (SIM), and composite interval mapping (CIM) can also be solved by an EM algorithm, as derived in (Kao and Zeng 1997).

Assuming the CIM model and iid Normal for the error term, the joint likelihood function (Kao and Zeng 1997) for $\boldsymbol{\theta} = (\mathbf{p}, a, d, \boldsymbol{\beta}, \sigma^2)$ of n individuals is as follows:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{3} p_{ji} \phi\left(\frac{y_j - \mu_{ji}}{\sigma} \right) \right]$$

where $\phi(\cdot)$ is a standard normal pdf, $\mu_{j1} = a - d/2 + X_j^*\boldsymbol{\beta}^*$, $\mu_{j2} = d/2 + X_j^*\boldsymbol{\beta}^*$, and $\mu_{j3} = -a - d/2 + X_j^*\boldsymbol{\beta}^*$. The index i iterates over genotypes AA, Aa, and aa. Thus, μ_{ji} and p_{ji} denote the mean and the prior probability (given flanking markers) of the i^{th} QTL genotype.

An EM algorithm can then be used to obtain maximum-likelihood estimates (MLE) of $\boldsymbol{\theta}$. The normal mixture of the preceding equation can be treated as an incomplete-data problem (Dempster et al. 1977) since the QTL genotypes are unknown. Let

$$g_j(x_j^*, z_j^*) = \begin{cases} p_{j1} & \text{if } x_j^* = 1 \text{ and } z_j^* = -\frac{1}{2} \\ p_{j2} & \text{if } x_j^* = 0 \text{ and } z_j^* = \frac{1}{2} \\ p_{j3} & \text{if } x_j^* = -1 \text{ and } z_j^* = -\frac{1}{2} \end{cases}$$

be the distribution of QTL genotype specified by x_j^* and z_j^*. The unobserved QTL genotypes, (x_j^* and z_j^*), are treated as missing data (Kao and Zeng 1997), denoted by q_j^*, and trait y_j,

selected background markers, and explanatory variables X_j are treated as observed data, denoted by $y_{(\text{obs},j)}$. The conditional distribution of observed data given missing data is

$$y_j|(\boldsymbol{\theta}, X_j, x_j^*, z_j^*) \sim N(x_j^* a + z_j^* d + X_j \boldsymbol{\beta}, \sigma^2)$$

Thus, the density of the complete-data, $y_{(com,j)}$, is

$$f(y_{(com,j)}|\boldsymbol{\theta}) = f(y_{(obs,j)}|\boldsymbol{\theta}, X_j, x_j^*, z_j^*) g(x_j^*, z_j^*)$$

The E step of the EM algorithm is as follows.

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \int \log L(\boldsymbol{\theta}|\mathbf{y}_{\text{com}}) f(\mathbf{q}^*|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{q}^* \\ &= \int \log \left[\prod_{j=1}^{n} \phi\left(\frac{y_j - \mu_j}{\sigma}\right) g_j(x_j^*, z_j^*) \right] \times f(\mathbf{q}^*|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{q}^* \end{aligned}$$

By Fubini's theorem (Fubini 1958):

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \sum_{j=1}^{n} \int \log \left[\phi\left(\frac{y_j - \mu_j}{\sigma}\right) g_j(x_j^*, z_j^*) \right] \times f(q_j^*|y_{(\text{obs},j)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, dq_j^* \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3} \log \left[\phi\left(\frac{y_j - \mu_j}{\sigma}\right) p_{ji} \right] \times \frac{p_{ji} \phi\left(\frac{y_j - \mu_{ji}^{(t)}}{\sigma^{(t)}}\right)}{\sum_{k=1}^{3} p_{jk} \phi\left(\frac{y_j - \mu_{jk}^{(t)}}{\sigma^{(t)}}\right)} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3} \log \left[\phi\left(\frac{y_j - \mu_j}{\sigma}\right) p_{ji} \right] \times \pi_{ji}^{(t)} \end{aligned}$$

Observe that by Bayes' rule π_{ji} is the posterior probability of the QTL genotype. Assuming that the true QTL genotype is the i^{th} genotype, write

$$\begin{aligned} f\left(q_j^*|y_{(\text{obs},j)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right) &= \frac{f\left(y_{(\text{obs},j)}|q_j^*, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right) f\left(q_j^*|\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right)}{f\left(y_{(\text{obs},j)}|\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right)} \\ &= \frac{\phi\left(\frac{y_j - \mu_{ji}^{(t)}}{\sigma^{(t)}}\right) p_{ji}}{\sum_{k=1}^{3} p_{jk} \phi\left(\frac{y_j - \mu_{jk}^{(t)}}{\sigma^{(t)}}\right)} = \pi_{ji} \end{aligned}$$

In the M step, the MLE solution is obtained by differentiation of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$ with respect to $\boldsymbol{\theta}$, yielding

$$\begin{aligned}
a^{(t+1)} &= \frac{\sum_{j=1}^{n}\left[\left(\pi_{j1}^{(t)}-\pi_{j3}^{(t)}\right)\left(y_j - X_j\boldsymbol{\beta}^{(t)}\right) - \frac{1}{2}\left(\pi_{j3}^{(t)}-\pi_{j1}^{(t)}\right)d^{(t)}\right]}{\sum_{j=1}^{n}\left(\pi_{j1}^{(t)}+\pi_{j3}^{(t)}\right)} \\
&= \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_1 - \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_2)d^{(t)}}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1)} \\
d^{(t+1)} &= \frac{\sum_{j=1}^{n}\frac{1}{2}\left[\left(-\pi_{j1}^{(t)}+\pi_{j2}^{(t)}-\pi_{j3}^{(t)}\right)\left(y_j - X_j\boldsymbol{\beta}^{(t)}\right) - \left(\pi_{j3}^{(t)}-\pi_{j1}^{(t)}\right)a^{(t)}\right]}{\sum_{j=1}^{n}\left(\pi_{j1}^{(t)}+\pi_{j2}^{(t)}+\pi_{j3}^{(t)}\right)} \\
&= \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_2 - \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_2)a^{(t)}}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2)}
\end{aligned}$$

where # denotes the Hadamard (componentwise) product of two vectors, $\boldsymbol{\Pi} = \{\pi_{ji}\}_{n\times 3}$, $\mathbf{d}_1 = (1,0,-1)'$, and $\mathbf{d}_2 = (-\frac{1}{2},\frac{1}{2},-\frac{1}{2})'$.

Let $\mathbf{e}^{(t)} = (a^{(t)}, d^{(t)})'$. The equations above simplify to

$$\mathbf{e}^{(t+1)} = \mathbf{r}^{(t)} - \mathbf{M}^{(t)}\mathbf{e}^{(t)}$$

where

$$\mathbf{r}^{(t)} = \begin{bmatrix} \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_1}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1)} \\ \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_2}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2)} \end{bmatrix} \quad \text{and} \quad \mathbf{M}^{(t)} = \begin{bmatrix} 0 & \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_2)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1)} \\ \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_1)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2)} & 0 \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{y} - \boldsymbol{\Pi}^{(t)}\mathbf{D}\mathbf{e}^{(t+1)}\right] \\
\sigma^{2^{(t+1)}} &= \frac{1}{n}\left[\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t+1)}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t+1)}\right) - \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t+1)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{D}\mathbf{e}^{(t+1)}\right. \\
&\quad \left. + \mathbf{e}'^{(t+1)}\mathbf{V}^{(t)}\mathbf{e}^{(t+1)}\right]
\end{aligned}$$

where $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2)$ and

$$\mathbf{V}^{(t)} = \begin{bmatrix} \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1) & \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_2) \\ \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_1) & \mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2) \end{bmatrix}$$

Note that $\mathbf{V}$, but not $\mathbf{M}$, is symmetric.

In the first iteration, $\mathbf{\Pi}$ is filled with the genotypic distribution obtained from the Markov-chain approach described previously. The vectors $\boldsymbol{\beta}$ and $\mathbf{e}$ are filled with the estimates from the least-squares method.

The estimates of the parameters and the $\mathbf{\Pi}$ matrix are updated as described until convergence. The null hypothesis is $a = d = 0$ and the log likelihood for the null hypothesis is also computed. A likelihood ratio score (LR) is computed and LOD score is obtained by $\text{LOD} = \text{LR}/(2 \log 10)$.

2.4 Multiple interval mapping (MIM)

MIM (Kao et al. 1999) builds upon the EM solution of composite interval mapping (CIM). Instead of fitting background markers, it fits q QTLs simultaneously. Thus, the index i in the EM solution above runs from 1 to 3^q (or 2^q in the absence of a dominance effect), accounting for all possible QTL genotype combinations. The joint likelihood function now becomes:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{3^q} p_{ji} \phi \left(\frac{y_j - \mu_{ji}}{\sigma} \right) \right]$$

The updating rule for the posterior probability of QTL genotype becomes:

$$\pi_{ji}^{(t+1)} = \frac{\phi \left(\frac{y_j - \mu_{ji}^{(t)}}{\sigma^{(t)}} \right) p_{ji}}{\sum_{k=1}^{3^q} p_{jk} \phi \left(\frac{y_j - \mu_{jk}^{(t)}}{\sigma^{(t)}} \right)} \qquad \forall i \in \{1, \ldots, 3^q\}$$

The genetic design matrix $\mathbf{D}_{3^q \times 2} = \mathbf{1}_q \otimes \{\mathbf{d}_1, \mathbf{d}_2\}$, if both additive and dominance effects are present, or $\mathbf{D}_{3^q \times 1} = \mathbf{1}_q \otimes \{\mathbf{d}_1\}$, if only an additive effect is present. The symbol $\otimes$ denotes the Kronecker product.

Consequently:

$$\mathbf{\Pi} = \{\pi_{ji}\}_{n \times 3^q} \qquad \mathbf{V} = \{\mathbf{1}' \mathbf{\Pi} (\mathbf{D}_i \# \mathbf{D}_j)\}_{e \times e}$$

$$\mathbf{r} = \left\{ \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_1}{\mathbf{1}'\boldsymbol{\Pi}(\mathbf{D}_i \# \mathbf{D}_i)} \right\}_{e \times 1} \qquad \mathbf{M} = \left\{ \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_i \# \mathbf{d}_j)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_i \# \mathbf{d}_i)} \right\}_{e \times e}$$

where e is the number of columns of $\mathbf{D}$.

The formulas above are a straightforward extension to allow estimation of both dominance effect and non-genetic factors. The original paper describes only the formulas for designs with only additive effect present and no non-genetic factors. Here, matrix $\mathbf{X}$ contains only non-genetic factors, while the other matrices and vectors are kept the same as those of CIM. If there are no non-genetic factors, then $\mathbf{X} = \mathbf{1}$.

Since fitting too many QTL at once is computationally expensive, MIM uses a stepwise or "chunkwise" selection method to add QTLs into the model. In stepwise selection, QTL is added to the model one by one. In chunkwise selection, several QTLs are added to the model at a time. The QTL selection proceeds just like usual model-selection procedures.

2.5 Multiple-trait composite interval mapping (MT-CIM)

Multiple-trait QTL analysis is QTL analysis applied to several traits simultaneously. Such analysis was shown (Jiang and Zeng 1995) to improve the statistical power of QTL detection test and the precision of parameter estimation by taking into account the correlation structure among the traits. In addition, such analysis provides formal procedures to test for pleiotropy, *i.e.*, whether a QTL affects all selected traits, and QTL-by-environment (Q×E) interaction.

Multiple-trait CIM (Jiang and Zeng 1995) builds upon the EM solution of single-trait CIM. The model is the same as that of CIM, except that all matrices and vectors are expanded to t variates. If t denotes the number of traits (*i.e.*, variates), the model is expressed by

$$\underset{n\times t}{\mathbf{Y}} = \underset{n\times 1}{\mathbf{x}}\,\underset{1\times t}{\mathbf{a}} + \underset{n\times 1}{\mathbf{z}}\,\underset{1\times t}{\mathbf{d}} + \underset{n\times(2k+p+1)}{\mathbf{X}}\,\underset{(2k+p+1)\times t}{\mathbf{B}} + \underset{n\times t}{\mathbf{E}}$$

where k is the number of cofactor markers (with two effects calculated per marker) and p is the number of non-genetic covariates. Notice that if $t = 1$, this model reduces to CIM's model.

Although the encoding of QTL genotypes is different, the steps to derive the solution are essentially the same. In the original paper, x_j takes values of 2, 1, and zero for respective QTL genotypes AA, Aa, and aa. z_j takes value 1 if the QTL genotype is Aa, otherwise 0.

The joint likelihood function of the data is

$$L_1 = \prod_{j=1}^{n} [p_{2j}\phi_2(\mathbf{y}_j) + p_{ij}\phi_1(\mathbf{y}_j) + p_{0j}\phi_0(\mathbf{y}_j)]$$

where the p_{ij} are the prior probability of QTL genotypes of AA, Aa, and aa and $\phi_i(\cdot)$ is multivariate normal with variance σ and mean $\mathbf{u}_{j2} = \mathbf{X}_j\mathbf{B} + 2\mathbf{a}$, $\mathbf{u}_{j1} = \mathbf{X}_j\mathbf{B} + \mathbf{a} + \mathbf{d}$, and $\mathbf{u}_{j0} = \mathbf{X}_j\mathbf{B}$, respectively.

The log-likelihood function is given by

$$\begin{aligned}
\ln(L_1) = k^* - \frac{n}{2}\ln\left|\hat{\mathbf{V}}\right| + \sum_{j=1}^{n} \ln\Bigg[& p_{2j}\exp\left(\frac{1}{2}[\mathbf{y}_j - 2\hat{\mathbf{a}} - \mathbf{X}_j\hat{\mathbf{B}}]\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - 2\hat{\mathbf{a}} - \mathbf{X}_j\hat{\mathbf{B}}]'\right) \\
& + p_{1j}\exp\left(\frac{1}{2}[\mathbf{y}_j - \hat{\mathbf{a}} - \hat{\mathbf{d}} - \mathbf{X}_j\hat{\mathbf{B}}]\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - \hat{\mathbf{a}} - \hat{\mathbf{d}} - \mathbf{X}_j\hat{\mathbf{B}}]'\right) \\
& + p_{0j}\exp\left(\frac{1}{2}[\mathbf{y}_j - \mathbf{X}_j\hat{\mathbf{B}}]\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - \mathbf{X}_j\hat{\mathbf{B}}]'\right)\Bigg] \\
= k^* - \frac{n}{2}\ln\left|\hat{\mathbf{V}}\right| & - \frac{1}{2}\sum_{j=1}^{n}[\mathbf{y}_j - \mathbf{X}_j\hat{\mathbf{B}}]\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - \mathbf{X}_j\hat{\mathbf{B}}]' \\
& + \sum_{j=1}^{n} \ln\Big(p_{2j}\exp[2\hat{\mathbf{a}}\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - \hat{\mathbf{a}} - \mathbf{X}_j\hat{\mathbf{B}}]'] \\
& + p_{1j}\exp[(\hat{\mathbf{a}} + \hat{\mathbf{d}})\hat{\mathbf{V}}^{-1}[\mathbf{y}_j - \frac{1}{2}\hat{\mathbf{a}} - \frac{1}{2}\hat{\mathbf{d}} - \mathbf{X}_j\hat{\mathbf{B}}]'] + p_{0j}\Big)
\end{aligned}$$

where $k^* = -nt\ln(2\pi)/2$ and $|\hat{\mathbf{V}}|$ is the determinant of the covariance matrix.

Differentiating the log-likelihood function with respect to its parameters yields

$$\begin{aligned}
\mathbf{a}^{(t+1)} &= \frac{\mathbf{q}_2^{'(t+1)}}{2\mathbf{q}_2^{'(t+1)}\mathbf{1}}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)}) \\
\mathbf{d}^{(t+1)} &= \left[\frac{\mathbf{q}_1^{'(t+1)}}{\mathbf{q}_1^{'(t+1)}\mathbf{1}} - \frac{\mathbf{q}_2^{'(t+1)}}{2\mathbf{q}_2^{'(t+1)}\mathbf{1}}\right](\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)}) \\
\mathbf{B}^{(t+1)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{Y} - (2\mathbf{q}_2^{(t+1)} + \mathbf{q}_1^{(t+1)})\mathbf{a}^{(t+1)} - \mathbf{q}_1^{(t+1)}\mathbf{d}^{(t+1)}] \\
\mathbf{V}^{(t+1)} &= \frac{1}{n}\Big[(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t+1)})'(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t+1)}) - 4(\mathbf{q}_2^{'(t+1)}\mathbf{1})\mathbf{a}^{'(t+1)}\mathbf{a}^{(t+1)} \\
&\quad -(\mathbf{q}_1^{'(t+1)}\mathbf{1})(\mathbf{a}^{(t+1)} + \mathbf{d}^{(t+1)})'(\mathbf{a}^{(t+1)} + \mathbf{d}^{(t+1)})\Big]
\end{aligned}$$

where $\mathbf{q}_2^{(t+1)}$ and $\mathbf{q}_1^{(t+1)}$ are the respective $n \times 1$ vectors of $q_{2j}^{(t+1)}$ and $q_{1j}^{(t+1)}$, and for $i = 0, 1, 2$,

$$q_{ij} = \frac{p_{ij}\phi_i^{(t)}(\mathbf{y}_j)}{\sum_{k=0}^{2} p_{kj}\phi_k^{(t)}(\mathbf{y}_j)}$$

There are several modes of hypothesis testing in multiple-trait analysis:

1. **Joint QTL mapping** Is there any QTL detected for any trait?

   Under joint-mapping the hypotheses to be tested are $H_0 : \mathbf{a} = \mathbf{d} = 0$ *vs.* H_A : otherwise. In this case, the log-likelihood under H_0 is given by

   $$\ln(L_0) = \ln\left[\prod_{j=1}^{n}\phi_0(\mathbf{y}_j)\right] = k - \frac{n}{2}\ln|\hat{\mathbf{V}}_\mathbf{0}| - nm/2$$

   where $\hat{\mathbf{V}}_\mathbf{0} = (\mathbf{Y} - \mathbf{X}\mathbf{B}_0)'(\mathbf{Y} - \mathbf{X}\mathbf{B}_0)/n$ and $\mathbf{B}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The LOD score is obtained from $\text{LOD} = -\frac{1}{\ln(10)}\ln\left(\frac{L_0}{L_1}\right)$.

2. **Test of pleiotropic effects** Is there any QTL that affects all traits?

   The hypotheses to be tested are $H_0 : a_i = 0$ or $d_i = 0$ for some trait i *vs.* $H_A : \mathbf{a} \neq 0$ and $\mathbf{d} \neq 0$ (*i.e.*, testing if the QTL effects are zero in at least one trait). In this case, there are multiple null hypotheses. The log-likelihood ratio score can be obtained from the formulas derived above, with the appropriate effects set to zero. Although the original paper (Jiang and Zeng 1995) did not explicitly mention a combined LOD

score, it is usually taken as the minimum of the LOD scores at a particular marker or interval.

3. **Test of close linkage versus pleiotropy** Is the detected pleiotropic QTL not an artefact of several closely-linked QTLs?

   A QTL that affects all traits is not necessarily a pleiotropic QTL. It may be an artefact of several closely linked QTLs. This test is designed to distinguish the two, especially if the pleiotropic LOD score peak is wide.

   Let pos(i) be the position of the currently tested QTL for trait i. The hypotheses to be tested are $H_0 : \text{pos}(i) = \text{pos}(j),\ \forall i, j \in \{1, \ldots, t\}$ *vs.* H_A : otherwise. In this case, the position of QTL at trait i is shifted a little bit (1–2 cM) to either side and tested with the QTL at trait j. A LOD score is then computed in similar manner. If H_0 is rejected, then the QTL involved is not a pleiotropic QTL.

4. **QTL by environment (Q×E) analysis** Does the QTL affect one trait differently from the others?

   This test is particularly useful when the traits being tested are the same trait (*e.g.*, grain yield) in the same set of individuals (*e.g.*, replicated lines) but measured in different environments.

   Let a_i and d_i be the additive and dominance effect of a given QTL for a given trait. The hypothesis to be tested is

   $$H_0 : (a_i = a_j = a) \text{ and } (d_i = d_j = d),\ \forall i, j \in \{1, \ldots, t\}$$

   vs. H_A : not H_0.

   Under H_0, the E step is similar to that of the full model except that a_i and d_i are

replaced by a and d. In the CM step

$$a^{(t+1)} = \frac{\mathbf{q}_2^{'(t+1)}}{2c^{(t)}\mathbf{q}_2^{'(t+1)}\mathbf{1}}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)})(\mathbf{V}^{(t)})^{-1}\mathbf{1}$$

$$d^{(t+1)} = \left[\frac{\mathbf{q}_1^{'(t+1)}}{c^{(t)}\mathbf{q}_1^{'(t+1)}\mathbf{1}} - \frac{\mathbf{q}_2^{'(t+1)}}{2c^{(t)}\mathbf{q}_2^{'(t+1)}\mathbf{1}}\right](\mathbf{Y} - \mathbf{X}\mathbf{B}^{(t)})(\mathbf{V}^{(t)})^{-1}\mathbf{1}$$

where $c^{(t)} = \mathbf{1}'\mathbf{V}^{(t)}\mathbf{1}$.

So, under H_0, the log-likelihood now becomes:

$$\begin{aligned}\ln(L_0) = k^* - \frac{n}{2}\ln|\hat{\mathbf{V}}| - \frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})' \\ + \sum_{j=1}^{n}\ln\Big[p_{2j}\exp[(2a)\mathbf{1}'\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{1}'a - \mathbf{x}_j\hat{\mathbf{B}})'] \\ + p_{1j}\exp[(a+d)\mathbf{1}'\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{1}'(a+d)/2 - \mathbf{x}_j\hat{\mathbf{B}})] + p_{0j}\Big]\end{aligned}$$

The LOD score is obtained as $\text{LOD} = -\frac{1}{\ln(10)}\ln\left(\frac{L_0}{L_1}\right)$.

2.6 Summary

Single-marker regression (SMR), and simple (SIM), composite (CIM) and multiple interval mapping (MIM) can be solved using the same general linear model framework. SMR and SIM use essentially the same model to detect QTL: the former uses marker data, while the latter uses interpolated QTL genotype estimates. CIM and MIM also use the same model. CIM uses marker data, while MIM uses the calculated QTL estimates as the background.

Multiple-trait CIM (MT-CIM) extends CIM for correlated traits (Jiang and Zeng 1995). MT-CIM exploits the correlation structure to improve the accuracy of QTL detection. In addition, it provides additional tests that cannot be performed for single traits, such as pleiotropy and QTL-by-environment interaction.

In the absence of multiple traits, MT-CIM reduces to single-trait CIM and its performance and accuracy are the same as that of CIM. Multiple-trait analyses, such as pleiotropy and QTL-by-environment tests, also can no longer be performed.

Although SIM and CIM are still widely used today, they have less power than, and have been largely superseded by, MIM (Kao et al. 1999). SIM and CIM may be useful for preliminary analyses because they can be computed rapidly. Although MIM computation is much slower than that of SMR, SIM, or CIM, modern computers can perform MIM computation in seconds to minutes.

Since MT-CIM improves upon CIM, multiple-trait extension of MIM (MT-MIM) should also improve upon MIM. Since MIM has been shown (Kao et al. 1999) to give narrower peaks for QTLs than CIM, MT-MIM is expected to be more precise than MT-CIM. This motivates the development of MT-MIM.

Chapter 3

Multiple-trait multiple interval mapping (MT-MIM)

3.1 Derivation

The MT-MIM solution is based on single-trait MIM (MIM) using the models of multiple-trait CIM. The model is as follows:

$$\underset{n\times t}{\mathbf{Y}} = \sum_{i=1}^{q} \left[\underset{n\times 1}{\mathbf{x}_i} \underset{1\times t}{\mathbf{a}_i} + \underset{n\times 1}{\mathbf{z}_i} \underset{1\times t}{\mathbf{d}_i} \right] + \underset{n\times(p+1)}{\mathbf{X}} \underset{(p+1)\times t}{\mathbf{B}} + \underset{n\times t}{\mathbf{E}}$$

where q is the number of QTLs being fitted simultaneously, t is the number of analyzed traits, p is the number of non-genetic fixed factors, and $\mathbf{E}$ is random error, iid Normal($\mathbf{0}$, $\mathbf{\Sigma}$).

The differences between MT-MIM and single-trait MIM or MT-CIM are as follows. Here, $\mathbf{Y}$ is a matrix, similar to that of MT-CIM, instead of a vector as in MIM. Instead of testing one QTL at a time in MT-CIM, the model tests q QTLs at a time, as in MIM. The reader will notice that the method development is very similar to that of MIM (Kao et al. 1999).

Using the same framework, the joint likelihood for the parameter $\boldsymbol{\theta} = (\mathbf{p}, \mathbf{a}, \mathbf{d}, \mathbf{B}, \mathbf{\Sigma})$ is

$$L_1 = L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{3^q} p_{ji} f\left(\mathbf{y}_j; \mu_{ji}, \mathbf{\Sigma}\right) \right]$$

where $f\left(\mathbf{y}_j; \mu_{ji}, \mathbf{\Sigma}\right)$ is the multivariate normal pdf with mean μ_{ji} and variance $\mathbf{\Sigma}$.

The E step of the EM algorithm is very similar to that of MT-CIM, except now the pdf used in the derivation is multivariate instead of univariate normal.

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \int \log L(\boldsymbol{\theta}|\mathbf{Y}_{\text{com}}) f(\mathbf{Q}^*|\mathbf{Y}_{\text{obs}}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{Q}^* \\ &= \int \log \left[\prod_{j=1}^{n} f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right) g_j(x_j^*, z_j^*) \right] \times f(\mathbf{Q}^*|\mathbf{Y}_{\text{obs}}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{Q}^* \end{aligned}$$

By Fubini's theorem (Fubini 1958):

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \sum_{j=1}^{n} \int \log \left[f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right) g_j(x_j^*, z_j^*) \right] \times f(q_j^*|y_{(\text{obs},j)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{q}_j^* \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3^q} \log \left[f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right) p_{ji} \right] \times \frac{p_{ji} f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right)}{\sum_{k=1}^{3^q} p_{jk} f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right)} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3^q} \log \left[f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right) p_{ji} \right] \times \pi_{ji}^{(t)} \end{aligned}$$

Note that the summation on index i is now from 1 to 3^q, as in MIM, instead of 3 as in MT-CIM.

Thus, the updating rule for the posterior probability of the QTL genotype becomes:

$$\pi_{ji}^{(t+1)} = \frac{f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right) p_{ji}}{\sum_{k=1}^{3^q} p_{jk} f\left(\mathbf{y}_j; \mu_{ji}, \boldsymbol{\Sigma}\right)} \qquad \forall i \in \{1, \ldots, 3^q\}$$

Note that the posterior probabilities of the q QTLs are updated at the same time using a multivariate instead of a univariate normal pdf.

Differentiate $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$ with respect to each parameter to yield

$$\mathbf{E}^{(t+1)} = \mathbf{r}^{(t)} - \mathbf{M}^{(t)}\mathbf{E}^{(t)}$$

where

$$\mathbf{r}^{(t)} = \begin{bmatrix} \frac{\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_1}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_1)} \\ \frac{\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_2}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2 \# \mathbf{d}_2)} \end{bmatrix} \quad \text{and} \quad \mathbf{M}^{(t)} = \begin{bmatrix} 0 & \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_2)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_1)} \\ \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2 \# \mathbf{d}_1)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2 \# \mathbf{d}_2)} & 0 \end{bmatrix}$$

Note that $\mathbf{E}$ is a $eq \times t$ matrix of effects (not of residuals), where e is the number of effects as defined in the previous chapter.

$$\mathbf{B}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{Y} - \mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)}\right]$$
$$\mathbf{\Sigma}^{(t+1)} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t+1)} - \mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)}\right)'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t+1)} - \mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)}\right)$$

where $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2)$, as defined in previous sections.

MT-MIM can be considered as a generalization of both MIM and MT-CIM. If $t = 1$, MT-MIM reduces to MIM, while $q = 1$, MT-MIM reduces to MT-CIM.

The hypothesis testing is very similar to that of multiple-trait CIM:

1. **Joint QTL mapping** Is there any QTL detected for any trait?

   Under joint-mapping the hypotheses to be tested are $H_0 : \mathbf{E} = 0$ *vs.* H_A otherwise. In this case, the log-likelihood under H_0 is given by

   $$\ln(L_0) = \ln\left[\prod_{j=1}^{n} f\left(\mathbf{y}_j; \mu_j, \mathbf{\Sigma}\right)\right] = k - \frac{n}{2}\ln|\hat{\mathbf{V}}_\mathbf{0}| - nm/2$$

   where $\hat{\mathbf{V}}_\mathbf{0} = (\mathbf{Y} - \mathbf{X}\mathbf{B}_0)'(\mathbf{Y} - \mathbf{X}\mathbf{B}_0)/n$ and $\mathbf{B}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The LOD score is obtained from $\text{LOD} = -\frac{1}{\ln(10)}\ln\left(\frac{L_0}{L_1}\right)$.

2. **Test of pleiotropic effects** Is there any QTL that affects all traits?

   The hypotheses to be tested are $H_A : \mathbf{E} \neq 0$ *vs.* H_0 otherwise (*i.e.*, the effects of at least one QTL are zero in at least one trait). In this case, there are multiple null hypotheses. The log-likelihood ratio score can be obtained from the formulas derived above, with the appropriate effects set to zero. The combined LOD score is defined as the minimum of the LOD scores at a particular interval.

3. **QTL by environment (Q×E) analysis** Does the QTL affect one trait differently from the others?

This test is useful when the traits being tested are the same trait (*e.g.* grain yield) in the same set of individuals (*e.g.*, replicated lines) but measured in different environments.

Let $\mathbf{a}_i$ and $\mathbf{d}_i$ be the additive and dominance effects of a given QTL for a given trait. The hypothesis to be tested is

$$H_0 : \mathbf{a}_i = \mathbf{a}_j = \mathbf{a} \text{ and } \mathbf{d}_i = \mathbf{d}_j = \mathbf{d}, \ \forall i, j \in \{1, \ldots, t\}$$

vs. H_A: not H_0.

$$\mathbf{a}^{(t+1)} = \mathbf{1}_t'(\boldsymbol{\Sigma}^{(t)})^{-1} \left[\frac{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)' \boldsymbol{\Pi}^{(t)}\mathbf{d}_1 - (\mathbf{1}_t' \otimes \mathbf{1}_n)'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_2)d^{(t)}}{c^{(t)}\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_1)} \right]$$

$$\mathbf{d}^{(t+1)} = \mathbf{1}_t'(\boldsymbol{\Sigma}^{(t)})^{-1} \left[\frac{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)' \boldsymbol{\Pi}^{(t)}\mathbf{d}_2 - (\mathbf{1}_t' \otimes \mathbf{1}_n)'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_2)a^{(t)}}{c^{(t)}\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2 \# \mathbf{d}_2)} \right]$$

where $c^{(t)} = \mathbf{1}_t'\boldsymbol{\Sigma}^{(t)}\mathbf{1}_t$ and $\mathbf{1}_t$ is a $t \times 1$ column vector of ones.

So, under H_0, the log-likelihood is obtained by substitution of $\mathbf{a}$ and $\mathbf{d}$ into L_1 formula. The LOD score is obtained similarly.

Although the original MIM may use either stepwise or chunkwise selection (Kao et al. 1999) to add QTLs into the model, my implementation uses stepwise selection. This is because it saves computation time while generally yielding the same model as the chunkwise method.

3.2 Implementation details

In general, MIM produces sharper peaks than CIM. Multiple-trait MIM (MT-MIM) is expected to behave similarly relative to MT-CIM. Since MIM produces sharper peaks than

CIM, the test of close linkage versus pleiotropy in MT-CIM is not useful in practice. For this reason, the implementation of this test is omitted in this thesis.

MT-CIM treats the variance–covariance matrix of the trait data as unstructured, which effectively limits the number of traits t because the number of estimated parameters in the matrix is $\frac{t(t+1)}{2}$. The MT-MIM method will also treat the matrix as unstructured, and hence, is subject to the same limitation.

Although the models used in the previously explained methods fall within the general linear model (GLM) framework, common statistical software does not necessarily provide solutions to all of them. SMR, SIM, and CIM models are merely simple or multiple regression, which can be solved by SAS or R. However, in MIM, a slight modification in the EM algorithm to accommodate QTL genotypic probabilities makes it necessary to move beyond standard linear-models packages.

Although the authors of MIM and multivariate CIM have published their software online, there are problems to using or extending them, prompting the need to develop custom software. The two programs are written in different languages: MIM in FORTRAN, multivariate CIM in C. In addition, their implementations are restricted to very few mating designs (only backcross and F_2), limiting their use and adoption by general analysts who are usually unfamiliar with programming and statistics.

QGene 4 (Joehanes and Nelson 2008), written in the Java language for multi-platform operation, accommodates rapid QTL-analysis method implementation with a plug-in architecture allowing component reusability. QGene 4 was a complete rewrite of an earlier application (Nelson 1997), which implemented only SMR and SIM. Current QGene 4 has CIM, MIM, and multiple-trait CIM implemented.

MT-MIM is built in QGene as a plug-in so that it can reuse many QGene components and hence shorten the development time. Moreover, QGene has a simulation module that can be used to compare the accuracy of MT-MIM and other methods.

Chapter 4

Results of QTL analysis using MT-MIM

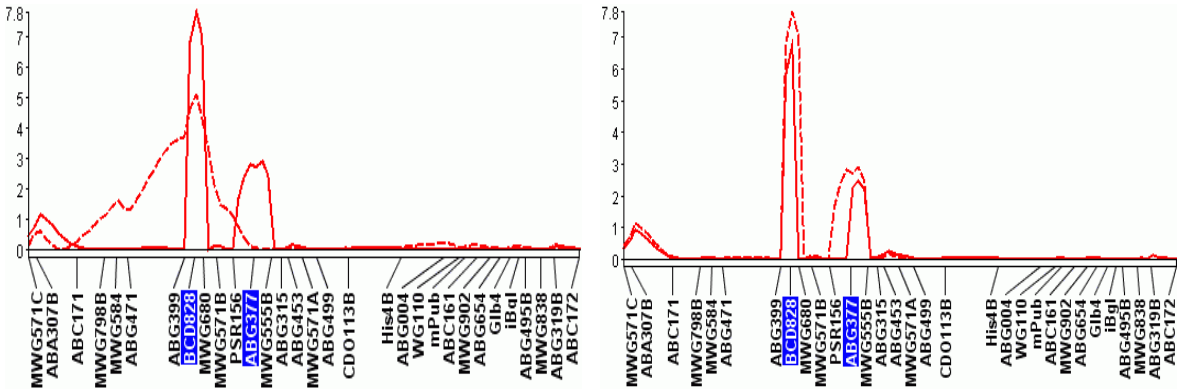
4.1 Description of the dataset and previous results

The data used in this experiment were derived from a cross between barley cultivars Steptoe and Morex (Kleinhofs et al. 1993), and contain grain yield data measured for 150 DH lines grown in 16 different environments (*yld01–yld16*). While there are no missing trait data, some genotype data are missing.

It is known that a QTL on chromosome 3 influences grain yield (Hayes et al. 1994). A plot of chromosome 3 shows profiles for simple (SIM), composite (CIM), and multiple interval mapping (MIM) is presented in Figure 4.1a and b. CIM gives narrower peaks than SIM. These figures confirm that CIM is more precise than SIM and MIM is more precise than CIM.

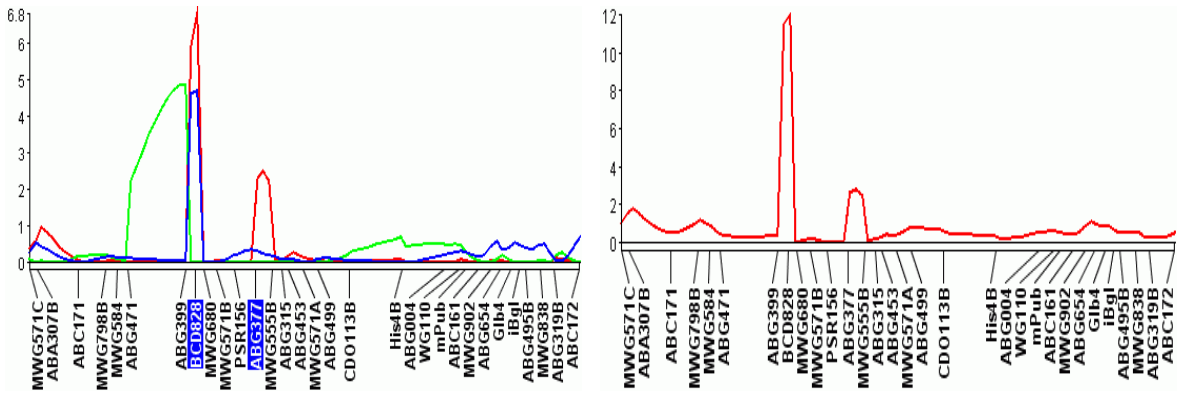
4.2 Results of MT-MIM on barley data

The preliminary results of MT-MIM applied to the first three traits (*yld01–yld03*) of the Steptoe × Morex barley data were promising. Not only does MT-MIM show a larger LOD score, MT-MIM also manages to "merge" disparate QTL locations on chromosome 3 as shown in Figure 4.2. Single-trait analysis typically cannot consolidate disparate QTL positions.



(a) (b)

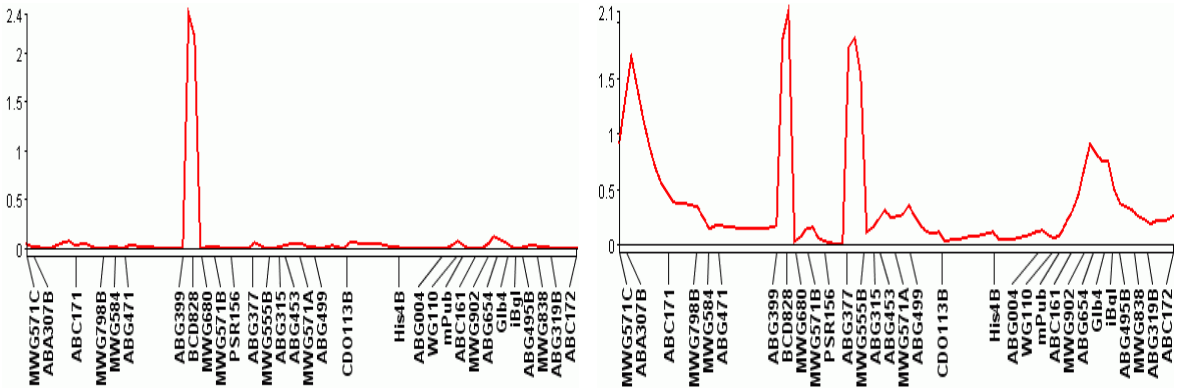
Figure 4.1: *(a) CIM improves upon SIM by giving narrower peaks. Broken line: SIM, solid line: CIM; (b) MIM improves upon CIM by giving even narrower peaks. Broken line: CIM, solid line: MIM. Highlighted markers represent cofactors selected for CIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*



(a) (b)

Figure 4.2: *(a) Single-trait MIM on traits* yld01–yld03 *gives different QTL location estimates in chromosome 3; (b) MT-MIM consolidates QTL locations and shows higher LOD scores. The QTL on chromosome 3 influences at least one of the three traits. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

The QTL in chromosome 3 is known to be pleiotropic and to act differently in different locations. MT-MIM is able to detect it, as shown in Figure 4.3.



(a) (b)

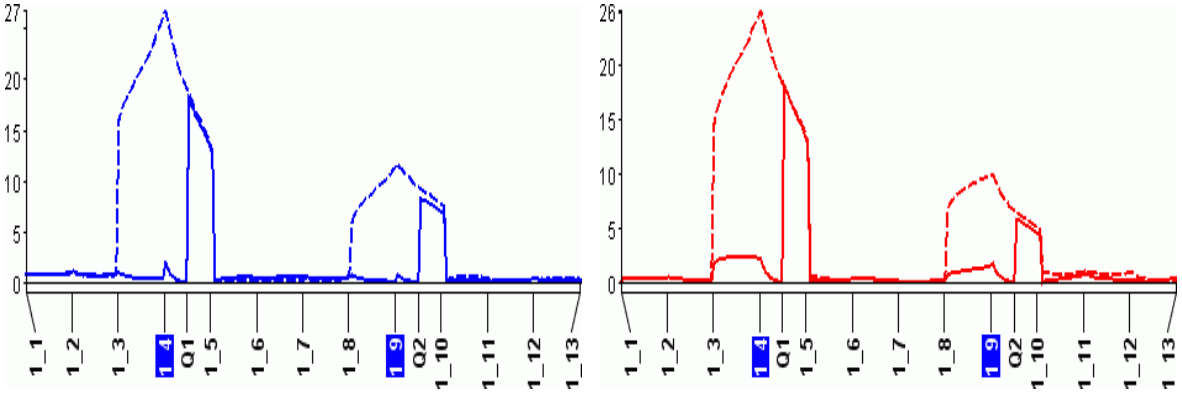
Figure 4.3: *(a) QTL on chromosome 3 is shown to be pleiotropic for traits* yld01–yld03. *This QTL influences all three traits.; (b) Q×E analysis of* yld01–yld03. *The QTL between markers* ABG399 *and* BCD828 *is seen to act differently in the three different environments. The QTL between markers* ABG377 *and* MWG555B *affects only trait* yld01. *The Q×E test correctly shows that the QTL acts differently in different environments. The regions on the left and right of the plot also show differences in different environments although show no QTLs. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

4.3 Simulation setup and results

One hundred simulated datasets of 200 progeny each were generated by QGene for two correlated traits in the F_2 mating design. The *heritability*, or the proportion of the phenotypic variance accounted for by genotypic variance, was set to 0.5. The trait correlations were chosen to be 0.9, 0.4, and 0.0, so that there were 300 datasets in total. For each trait there were two QTLs in the model, Q_1 at 35 centiMorgans (cM) and Q_2 at 85 cM, with the effects described in Table 4.1. The QTLs were placed on a 120-cM chromosome, with markers at 10-cM intervals. Scan interval was set to 1 cM. A QTL was declared if there was a LOD score of at least 2.8 within 15 cM of the true QTL position.

In the presence of correlation, multiple-trait CIM (MT-CIM) produced a wider peak than multiple-trait MIM (MT-MIM) in all datasets. A typical example is shown in Figure

4.4. Both methods detected all QTLs in the data 97% of datasets.



(a) (b)

Figure 4.4: *Comparison between MT-MIM (solid line) and MT-CIM (broken line) when the correlation between the two traits is (a) $\rho = 0.4$; (b) $\rho = 0.9$. Highlighted markers represent cofactors selected for MT-CIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

In the absence of correlation, MT-CIM produced sharper peaks than those in the presence of correlation, but MT-MIM produced even sharper peaks, as shown in Figure 4.5. Both methods could detect all QTLs in the data.

QTLs detected by single-trait MIM tended to differ for each trait, regardless of the presence of correlation. A typical example is shown in Figures 4.6 and 4.7. Although single-trait MIM tended to either narrowly miss the QTL (25% of all cases) or completely fail to detect weaker QTL Q_2 (30.4% of all cases), MT-MIM could detect all QTLs correctly.

In all cases, MT-CIM produced higher and wider peaks than MT-MIM, even after co-factors next to the QTLs were selected. Although the peaks produced by MT-MIM show lower LOD scores, they passed the permutation analysis thresholds. The average signifi-

Table 4.1: *QTL effect values from comparative simulation study of MT-MIM*

Trait	**Q_1 effect**		**Q_2 effect**	
	Additive	**Dominance**	**Additive**	**Dominance**
Trait 1	1.0	0.0	0.3	0.0
Trait 2	0.7	0.0	0.4	0.0

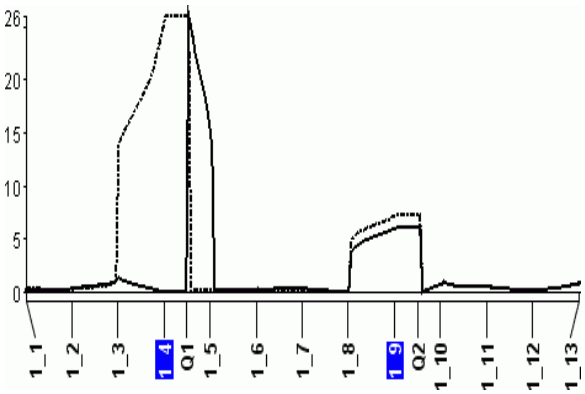
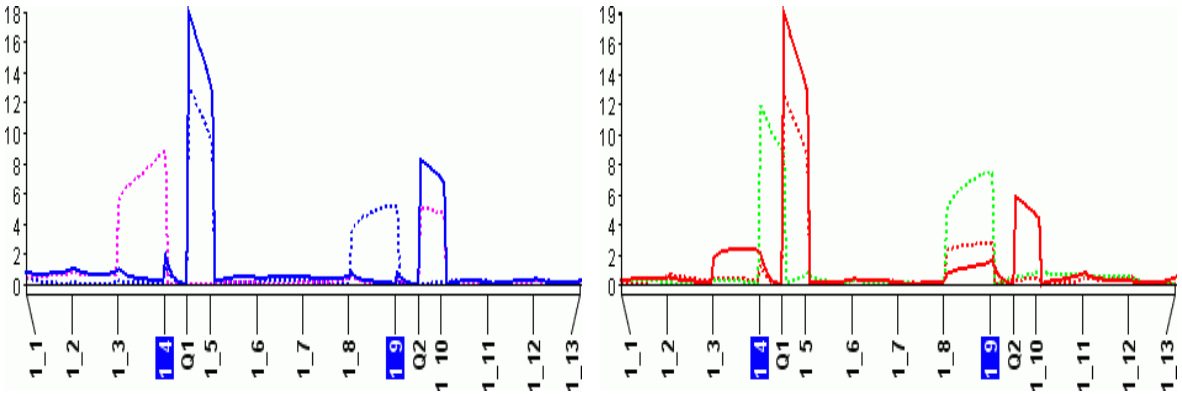


Figure 4.5: *Comparison between MT-MIM (solid line) and MT-CIM (broken line) when there is no correlation between the traits. Highlighted markers represent cofactors selected for MT-CIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*



(a) (b)

Figure 4.6: *Comparison between MT-MIM (solid line) and single-trait MIM (broken line) for two traits with correlation (a) $\rho = 0.4$; (b) $\rho = 0.9$ and controlled by the same two QTLs. Blue and red lines: trait 1; magenta and green lines: trait 2. These figures show that single-trait MIM may give inconsistent QTL location estimates. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

cance thresholds for MT-MIM of all datasets were 3.70 and 4.59, for $\alpha = 0.05$ and 0.01. For MT-CIM, they were 3.31 and 4.3.

In single-trait MIM, the low LOD score at QTL Q_2 failed to pass the significance threshold. For single-trait MIM, the average significance thresholds were 2.66 and 3.48.

For comparison, single-trait Bayesian interval mapping (BIM) was run on one of the datasets, as shown in Figure 4.8(a). The result suggests that unlike single-trait MIM, BIM does not suffer from QTL location-estimate disagreements. However, BIM requires much longer computation time than MIM. In addition, the computation of the QTL significance threshold using permutation takes a prohibitively long time. One 1,000-iteration permutation run may take 3 to 4 hours on a dual-core Intel Core 2 Duo T9300, 2.5GHz laptop, with parallel algorithms enabled.

Simple interval mapping (SIM) performed very poorly, as expected. SIM produced a very wide peak and failed to distinguish QTLs Q_1 and Q_2, as shown in Figure 4.8(b).

The simulation results are summarized in Table 4.2.

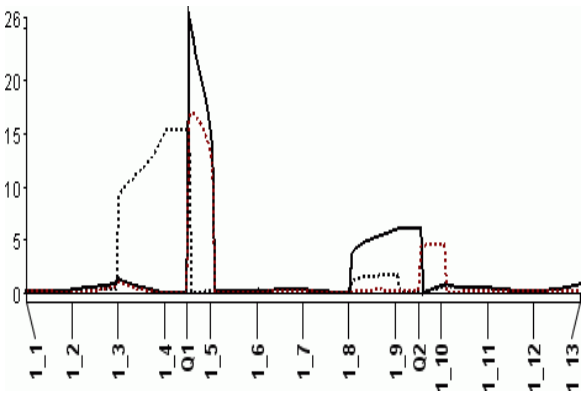
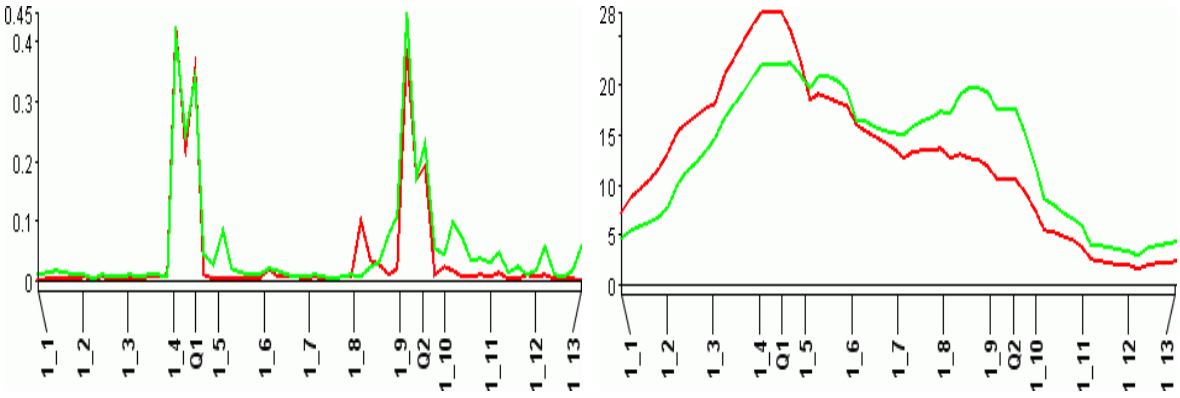


Figure 4.7: *Comparison between MT-MIM (solid line) and single-trait MIM (broken line) for two uncorrelated traits controlled by the same two QTLs. Black broken line: trait 1; brown broken line: trait 2. This figure shows that single-trait MIM may give inconsistent QTL location estimates for two traits controlled by the same QTLs even when they are uncorrelated. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*



(a) (b)

Figure 4.8: *(a) Single-trait Bayesian interval mapping (BIM) shows consistent QTL location estimates for two traits controlled by the same QTLs. (b) Simple interval mapping (SIM) fails to distinguish the two QTLs. Red line: trait 1; green line: trait 2. Vertical axis: for BIM, QTL posterior probability; for SIM, LOD score. Horizontal axis represents genetic location on the chromosome.*

Table 4.2: *Comparative QTL-detection accuracy of three QTL analysis methods for two traits controlled by the same QTLs, from simulation. Entries denote the percentage of the corresponding QTLs detected by the methods.*

Method	$\rho = 0.0$		$\rho = 0.4$		$\rho = 0.9$	
	$\mathbf{Q}_1$	$\mathbf{Q}_2$	$\mathbf{Q}_1$	$\mathbf{Q}_2$	$\mathbf{Q}_1$	$\mathbf{Q}_2$
MT-CIM	1.000	0.968	1.000	0.971	1.000	0.978
MIM (trait 1)	1.000	0.425	1.000	0.435	1.000	0.396
MIM (trait 2)	1.000	0.977	1.000	0.966	1.000	0.976
MT-MIM	1.000	0.977	1.000	0.973	1.000	0.982

Chapter 5

Discussion and conclusion

In the simulation study presented in the preceding chapter, it was shown that when the traits were governed by a set of common QTLs, single-trait analyses sometimes failed to agree on the QTL location estimates. Multiple-trait methods tended to merge the disparate QTL location estimates, even in the absence of trait correlation.

The study also showed that MT-MIM improves upon MT-CIM by giving narrower QTL peaks. Narrow peaks mean a smaller space for subsequent experimental searches for the QTL.

Why can multiple-trait analyses merge the disparate peaks given by single-trait analyses? To simplify the argument, consider a bivariate regression setting with only additive effects, with a putative QTL k. Let y_1 and y_2 be the trait values and x_k be the encoded genotype of QTL k. Let $\mathbf{b}_k = (b_{k1}, b_{k2})$ be the regression coefficients of y_1 and y_2 on x_k. Let σ_i^2 and ρ be the residual variance for trait i and the residual correlation coefficient. Let $\sigma_{x_k}^2$ be the variance of x_k. Let $\beta_i = b_{ki}\sigma_{x_k}/\sigma_i$. (Jiang and Zeng 1995) showed in their Appendix A that the likelihood ratio test statistic for the joint analysis ($\mathrm{LR}_{\mathrm{joint}}$) is expected to be:

$$\begin{aligned}\mathrm{LR}_{\mathrm{joint}} &= n \ln\left[1 + \frac{\beta_1^2 + \beta_2^2 - 2\rho\beta_1\beta_2}{1-\rho^2}\right] \\ &\simeq n\frac{\beta_1^2 + \beta_2^2 - 2\rho\beta_1\beta_2}{1-\rho^2}, \qquad \text{if } \beta_1^2, \beta_2^2 \ll 1.\end{aligned}$$

The $\mathrm{LR}_{\mathrm{joint}}$ is effectively a function of the sum of the QTL effect of each trait weighted by the inverse of the trait residual variance. Thus, the effect and location of the QTL that is

detected on the trait with lower residual variance will dominate and "unify" the effect and location of QTLs on the other traits, even if $\rho = 0$. Let $\mathbf{I}_t$ be a $t \times t$ identity matrix and $\mathbf{V}$ be the variance–covariance matrix. Since $L_0 \propto |\mathbf{V} + \sigma_k^2 \mathbf{b}_k' \mathbf{b}_k|^{-n/2}$ and $L_1 \propto |\mathbf{V}|^{-n/2}$, the general formula of expected $\mathrm{LR}_{\mathrm{joint}}$ is:

$$\begin{aligned} \mathrm{LR}_{\mathrm{joint}} &= -2 \ln(L_0/L_1) \\ &= n \ln \left(\frac{|\mathbf{V} + \sigma_k^2 \mathbf{b}_k' \mathbf{b}_k|}{|\mathbf{V}|} \right) \\ &= n \ln \left| \mathbf{I}_t + \sigma_k^2 \mathbf{b}_k' \mathbf{b}_k \mathbf{V}^{-1} \right| \end{aligned}$$

In this formulation, $\mathbf{V}^{-1}$ serves as the weight matrix that aggregates the effects of the QTL across traits. Even if there is no correlation among the traits, *i.e.*, $\mathbf{V}$ is diagonal, the effects of other traits will still influence $\mathrm{LR}_{\mathrm{joint}}$ because the term $\mathbf{b}_k' \mathbf{b}_k$ will never be diagonal if there are at least two traits influenced by the same QTL. In this case, the QTL effect on the trait with lower residual variance will prevail. This argument is also applicable for cases of multiple markers and dominance effect. The derivation is similar to the one presented in (Jiang and Zeng 1995).

Although MT-MIM takes more computation time, it was still computed very rapidly in the simulation study. The computation time depends on the number of chromosomes, chromosome length, scan interval, number of observations, and number of traits analyzed at once. For six traits and 12 chromosomes of about 200 cM, MT-CIM may take a minute or two in modern computers.

Both MT-CIM and MT-MIM treat the variance–covariance matrix as unstructured, which effectively limits the number of traits t because the number of estimated parameters in the matrix is $\frac{t(t+1)}{2}$. The more traits included in the model, the closer the variance–covariance matrix approaches to singularity. Without having investigated the upper limit on the number of traits that should be included for a given number n of progeny, I suggest that it should not exceed $\lceil \frac{1}{2}\sqrt{n} \rceil$, where $\lceil x \rceil$ is the smallest integer greater than or equal to

x. Other analyses such as mixed linear models must be used for higher numbers of traits, as adding the number of observations required for reliable estimates becomes impractical.

Despite some limitations, MT-MIM shows potential as an emerging multiple-trait QTL analysis method. It combines the accuracy of MIM and the sensitivity of multiple-trait analysis with reasonable computation time. Further investigation could involve a more extensive simulation study to investigate more of MT-MIM's properties, especially in different mating designs, in the presence of dominance effects, and in cases of linked QTLs.

Bibliography

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

Benjamini, Y. and Yekutieli, D. (2005), "Quantitative trait loci analysis using false discovery rate," *Genetics*, 171, 783–790.

Churchill, G. A. and Doerge, R. W. (1994), "Empirical threshold values for quantitative trait mapping," *Genetics*, 138, 963–971.

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

Doerge, R. W. (1995), "The relationship between the LOD score and the analysis of variance F-statistic when detecting QTL using single markers. Appendix 1: locating genes associated with root morphology and drought avoidance in rice via linkage to molecular markers," *Theoretical and Applied Genetics*, 90, 969–981.

Fubini, G. (1958), "Sugli integrali multipli," *Cremonese*, 2, 243–249.

Haldane, J. B. S. (1919), "The combination of linkage values, and the calculation of distances between the loci of linked factors," *Journal of Genetics*, 8, 299–309.

Hayes, P., Liu, B., Knapp, S., Chen, F., Jones, B., Blake, T., Franckowiak, J., Rasmusson, D., Sorrels, M., Ullrich, S., Wesenberg, D., and Kleinhofs, A. (1994), "Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm," *Theoretical and Applied Genetics*, 87, 392–401.

Jansen, R. C. (1994), "Controlling the type I and type II errors in mapping quantitative trait loci," *Genetics*, 138, 871–881.

Jansen, R. C. and Stam, P. (1994), "High resolution of quantitative traits into multiple loci via interval mapping," *Genetics*, 136, 1447–1455.

Jiang, C. and Zeng, Z.-B. (1995), "Multiple trait analysis of genetic mapping for quantitative trait loci," *Genetics*, 140, 1111–1127.

Jiang, C. and Zeng, Z. B. (1997), "Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines," *Genetica*, 101, 47–58.

Joehanes, R. and Nelson, J. C. (2008), "QGene 4.0, an extensible Java QTL-analysis platform," *Bioinformatics*, 24, 2788–2789.

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

Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. (1999), "Multiple trait mapping for quantitative trait loci," *Genetics*, 152, 1203–1216.

Kleinhofs, A., Kilian, A., Maroof, M. S., Biyashev, R., Hayes, P., Chen, F., Lapitan, N., Fenwick, A., Blake, T., Kanazin, V., Ananiev, E., Dahleen, L., Kudrna, D., Bollinger, J., Knapp, S., Liu, B., Sorrells, M., Heun, M., Franckowiak, J., Hoffman, D., Skadsen, R., and Steffenson, B. (1993), "A molecular, isozyme, and morphological map of the barley (*Hordeum vulgare*) genome," *Theoretical and Applied Genetics*, 86, 705–712.

Lander, E. S. and Botstein, D. (1989), "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps," *Genetics*, 121, 185–199.

Mangin, B., Goffinet, B., and Rebai, A. (1994), "Confidence intervals for QTL location," *Genetics*, 138, 1301–1308.

Nelson, J. C. (1997), "QGENE: software for marker-based genomic analysis and breeding," *Molecular Breeding*, 3, 239–245.

Satagopan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996), "A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo," *Genetics*, 144, 805–816.

Sillanpää, M. J. and Arjas, E. (1998), "Bayesian mapping of multiple quantitative trait loci from incomplete data based on line crosses," *Genetics*, 148, 1373–1388.

Stephens, D. A. and Fisch, R. D. (1998), "Bayesian mapping of multiple quantitative trait loci from incomplete data based on line crosses," *Biometrics*, 54, 1334–1347.

Storey, J. and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Zeng, Z.-B. (1994), "Precision Mapping of Quantitative Trait Loci," *Genetics*, 136, 1457–1468.