# GENERALIZED AND MULTIPLE-TRAIT EXTENSIONS TO QUANTITATIVE-TRAIT LOCUS MAPPING

by

ROBY JOEHANES

B.S., Universitas Pelita Harapan, Indonesia, 1999
M.S., Kansas State University, 2002
M.S., Kansas State University, 2009

---

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics Program
Department of Plant Pathology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2009

# Abstract

QTL (quantitative-trait locus) analysis aims to locate and estimate the effects of genes that are responsible for quantitative traits, by means of statistical methods that evaluate the association of genetic variation with trait (phenotypic) variation. Quantitative traits are typically controlled by multiple genes with varying degrees of influence on the phenotype. I describe a new QTL analysis method based on shrinkage and a unifying framework based on the generalized linear model for non-normal data. I develop their extensions to multiple-trait QTL analysis. Expression QTL, or eQTL, analysis is QTL analysis applied to gene expression data to reveal the eQTLs controlling transcript-abundance variation, with the goal of elucidating gene regulatory networks. For exploiting eQTL data, I develop a novel extension of the graphical Gaussian model that produces an undirected graph of a gene regulatory network. To reduce the dimensionality, the extension constructs networks one cluster at a time. However, because Fuzzy-K, the clustering method of choice, relies on subjective visual cutoffs for cluster membership, I develop a bootstrap method to overcome this disadvantage. Finally, I describe QGene, an extensible QTL- and eQTL-analysis software platform written in Java and used for implementation of all analyses.

# GENERALIZED AND MULTIPLE-TRAIT EXTENSIONS TO QUANTITATIVE-TRAIT LOCUS MAPPING

by

ROBY JOEHANES

B.S., Universitas Pelita Harapan, Indonesia, 1999
M.S., Kansas State University, 2002
M.S., Kansas State University, 2009

---

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics Program
Department of Plant Pathology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2009

Approved by:

Major Professor
James C. Nelson

# Copyright

ROBY JOEHANES

2009

# Abstract

QTL (quantitative-trait locus) analysis aims to locate and estimate the effects of genes that are responsible for quantitative traits, by means of statistical methods that evaluate the association of genetic variation with trait (phenotypic) variation. Quantitative traits are typically controlled by multiple genes with varying degrees of influence on the phenotype. I describe a new QTL analysis method based on shrinkage and a unifying framework based on the generalized linear model for non-normal data. I develop their extensions to multiple-trait QTL analysis. Expression QTL, or eQTL, analysis is QTL analysis applied to gene expression data to reveal the eQTLs controlling transcript-abundance variation, with the goal of elucidating gene regulatory networks. For exploiting eQTL data, I develop a novel extension of the graphical Gaussian model that produces an undirected graph of a gene regulatory network. To reduce the dimensionality, the extension constructs networks one cluster at a time. However, because Fuzzy-K, the clustering method of choice, relies on subjective visual cutoffs for cluster membership, I develop a bootstrap method to overcome this disadvantage. Finally, I describe QGene, an extensible QTL- and eQTL-analysis software platform written in Java and used for implementation of all analyses.

# Table of Contents

# List of Figures

# List of Tables

# List of Program Listings

# Acknowledgements

I would like to thank my major professor, 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 my committee members, Drs. Jianming Yu, Doina Caragea, Peter Bradbury, and Lawrence Davis, for their input in this dissertation.

I would like to thank Drs. Barbara Valent and John Boyer for their approval to pursue a second degree in statistics.

# Dedication

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

# Chapter 1

# Introduction to quantitative trait locus analysis


## Abstract

In this chapter, the concept of quantitative trait locus (QTL) mapping and its related concepts are introduced. Basic terminology in QTL mapping is described and explained.


## 1.1 Background

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.

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

## 1.2 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.

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 chromosome pairs. 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 $Aa$.

## 1.3 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.

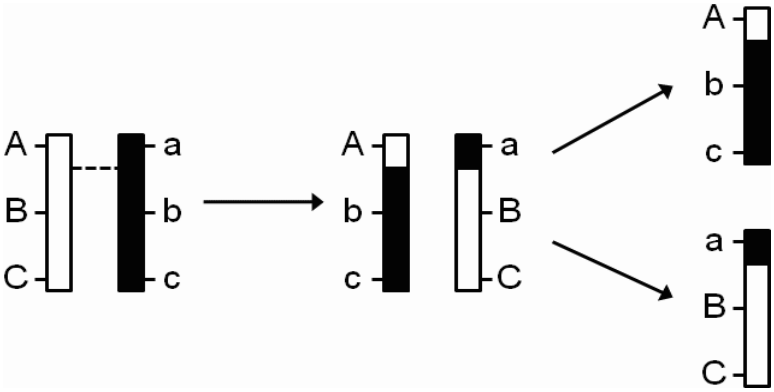



**Figure 1.1**: *A pair of chromosomes exchange segments via crossing over. The broken line shows the crossover point.*

## 1.4 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 $r_{AB} = (f_{Ab} + f_{aB})/(f_{AB} + f_{ab} + f_{Ab} + f_{aB})$.

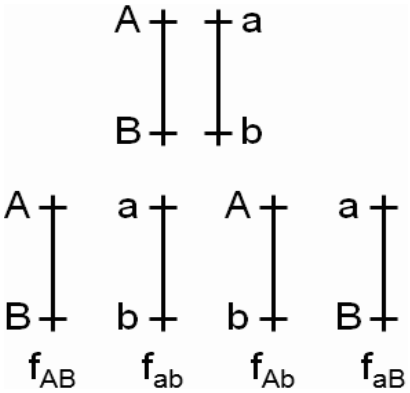



**Figure 1.2**: *The assortment of gametes produced in gametogenesis. The upper half shows the parental chromosome pair, with genotype AaBb; the lower half shows the array of gametes that it can produce. AB and ab represent parental and Ab and aB recombinant gametes. $f(XY)$ denotes the frequency of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

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 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.

## 1.5 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*.

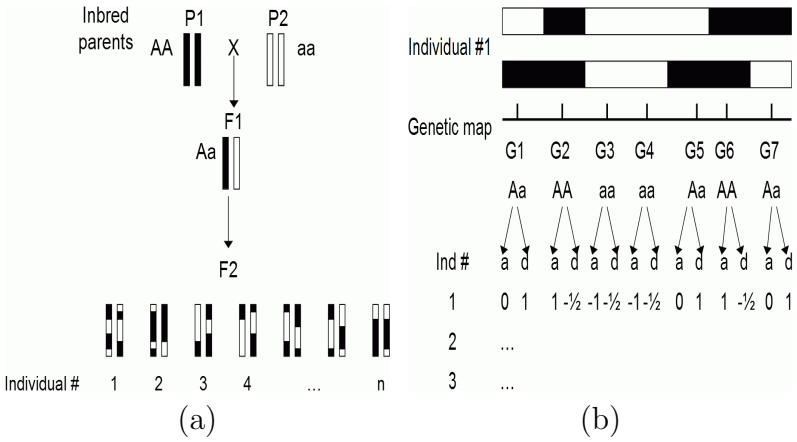



**Figure 1.3**: *(a) $F_2$ mating design; (b) Converting genotype data to numerical values*

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.

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 $F_2$ 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.

## 1.6 Genotypic assay and numerical conversion

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 half the difference between the phenotypic means (for some trait) 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.

## 1.7 Summary of a QTL mapping process

In summary, QTL mapping is done as follows:

- Select two inbred lines that have contrasting values for 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.

**Table 1.1**: *Table for converting genotypes into numerical values 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 combination of genotypes Aa and aA in one class.*

| **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 | | |

- 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 QTL mapping methods


## Abstract

In this chapter, most well-known QTL mapping methods, such as single-marker regression, simple, composite, multiple, multiple-trait composite, and Bayesian interval mapping, are compared and described in depth. The concept of inferring QTL genotypes is also explained.


## 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 for sufficient odds of detection of a QTL, and precisions 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 statistical-significance threshold for QTLs. The conventional threshold of LOD 3 based on the asymptotic distribution may not 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 (proportion of 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* a marker interval 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 use-

ful 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 Morgans. 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 from 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 $3 \times 1$ 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 $3 \times 1$ 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 such as 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.

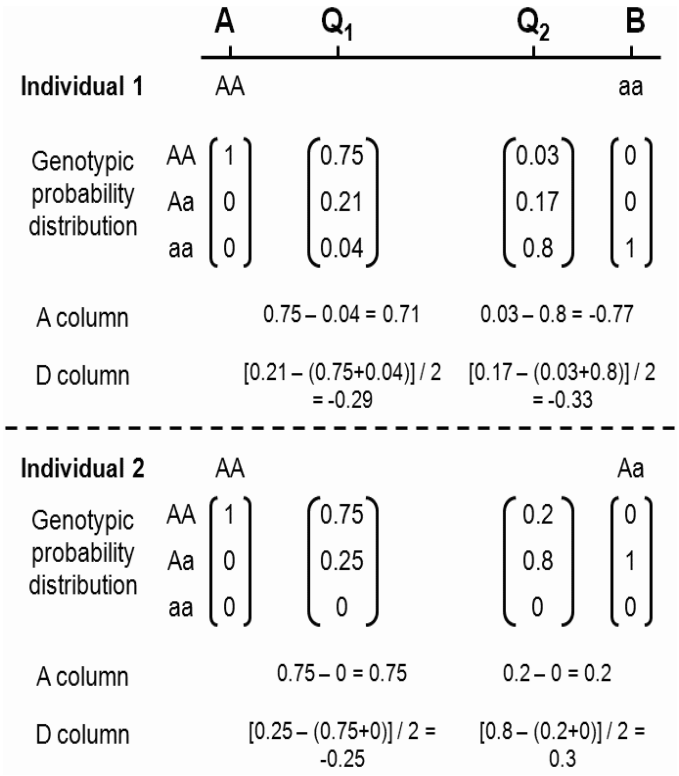



**Figure 2.1**: *An illustration of a Markov-chain method of (*Jiang and Zeng *1997) for inferring QTL genotype. In this example, QTL $Q_1$ and $Q_2$ are flanked by markers A and B. The genotypic probability distribution (GPD) vectors of $Q_1$ and $Q_2$ are computed with a Markov-chain method. Intuitively, 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.*

## 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 $\mu$ 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 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

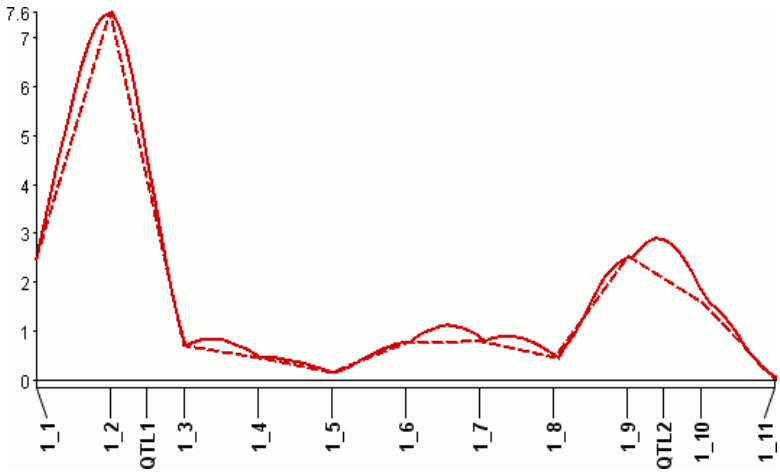



**Figure 2.2**: *Typical QTL-mapping profile. In this example, both single-marker regression (SMR) and simple interval mapping (SIM) detect a QTL 1, but only SIM can detect QTL 2. The plot is created from the LOD profiles of a simulated 200-progeny $F_2$ population. Broken line: SMR, solid line: SIM.*

design) and $m$ is the number of background markers. These can be selected manually or with 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 regression models and 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, these models can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. 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.*, $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$) 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. The LOD score can be obtained as 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, $\mathrm{df}_R$ and $\mathrm{df}_E$ 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, $\mu_{ji}$ and $p_{ji}$ denote the mean and the prior probability (given flanking markers) of the $i^{\text{th}}$ 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 governing changes to the order of integration,

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

Observe that by Bayes' rule, $\pi_{ji}$ is the posterior probability of the QTL genotype. Assuming that the true QTL genotype is the $i^{\text{th}}$ 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 the LOD score is obtained by $\text{LOD} = \text{LR}/(2\log 10)$.

## 2.4 MIM

Multiple interval mapping (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 as in usual model-selection procedures.

## 2.5 MT-CIM

Multiple-trait QTL analysis is QTL analysis applied to several traits simultaneously. Such analysis has been 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 (MT-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 the CIM 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 $\sigma$ 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

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

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 for trait $i$ is shifted a little bit (1–2 cM) to either side and tested with the QTL for trait $j$. A LOD score is then computed in a 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 for 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 BIM

In Bayesian interval mapping (BIM), Bayes' rule, instead of least squares or EM algorithm, is used to derive the solution, although the model used is the same as in CIM. Bayesian solutions require the posterior density of each parameter, obtained from the application of Bayes' rule. Samples are then generated from this density to yield parameter estimates. If the density resolves to a known one, the samples can be easily generated. However, it often does not. In this case, a sampling method based on Markov chain Monte Carlo (MCMC) approach is used to draw samples from approximate distributions of the parameters and use the samples to correct subsequent draws to better approximate the target posterior distribution. In MCMC, the samples are drawn sequentially, with the distribution of the sampled draws depending on the preceding values (states) drawn; hence, the draws form a Markov chain. Under the detailed-balance property, *i.e.,* where any given state of the chain

can be revisited infinite times from any state in a finite number of steps, the approximate distributions are improved at each step and converge to the target distribution (Gelman *et al.* 2003, pp. 285–286).

Gibbs sampling is an MCMC-based method in which samples of a parameter $\theta_j$ are drawn from the conditional distribution of $\theta_j$ given the values of all other parameters, $\theta_{-j}$, given data, $\mathbf{y}$, $p(\theta_j|\theta_{-j}, \mathbf{y})$. This means that Gibbs sampling requires $p(\theta_j|\theta_{-j}, \mathbf{y})$ for each parameter $j$ in the model. For many problems involving standard statistics, this conditional distribution resolves to a known form that is easily sampled.

For cases where conditional distributions are unknown, the Metropolis–Hastings (MH) algorithm is used. In MH, a vector of proposal values $\boldsymbol{\theta}^*$ is sampled at iteration $t$ for $\boldsymbol{\theta}$ from a proposal distribution of $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$. The proposal $\boldsymbol{\theta}^*$ is accepted (*i.e.*, $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$) at rate $\min(1, r)$, where:

$$r = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})/J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}{p(\boldsymbol{\theta}^{t-1}|\mathbf{y})/J_t(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}$$

and $\boldsymbol{\theta}^{t-1}$ is the values of the parameter vector at the previous iteration. In essence, the MH algorithm is an adaptation of a random walk that uses an acceptance–rejection rule to converge to the specified target distribution. It has been shown that Gibbs sampling is a special case of the MH algorithm with $r = 1$ (Gelman *et al.* 2003, pp. 292–293).

QTL mapping may be approached with a hybrid MH–Gibbs method (Satagopan *et al.* 1996). QTL locations and genotypes are sampled with Gibbs samplers, while QTL effects are sampled with the MH algorithm.

The MCMC method approach discussed so far assumes that the dimension of parameter space $\boldsymbol{\theta}$ is known. In QTL analysis, this implies that the number of QTLs is already known, which is rarely the case. For addressing this variable dimensionality, reversible jump MCMC (RJMCMC), a generalization of the MH algorithm allowing random walk across dimensions, was developed (Green 1995). It relies on a one-to-one, deterministic, and differentiable function $f_{n\to m}(\cdot)$ that transforms the parameter from dimension $n$ to $m$. Let $J_{n\to m}(\cdot|n, \boldsymbol{\theta}_n)$ be the proposal distribution that proposes parameter values in dimension $m$ given the values

of parameters in dimension $n$, $q(m|n)$ the probability of jumping from dimension $n$ to $m$, $p(n, \boldsymbol{\theta}_n|\mathbf{y})$ the posterior probability of parameter $\boldsymbol{\theta}_n$ in dimension $n$, $\mathcal{J}_{f_{n\to m}}$ the Jacobian term specified by the transformation function $f_{n\to m}(\cdot)$. The move from dimension $n$ to $m$ is accepted (*i.e.*, $\boldsymbol{\theta}^t = \boldsymbol{\theta}_m$) at a rate of $\min(1, r)$, where:

$$r = \frac{p(m, \boldsymbol{\theta}_m|\mathbf{y})/[q(m|n)J_{n\to m}(u_{n,m}|n, \boldsymbol{\theta}_n)]}{p(n, \boldsymbol{\theta}_n|\mathbf{y})/[q(n|m)J_{m\to n}(u_{m,n}|m, \boldsymbol{\theta}_m)]} \times \mathcal{J}_{f_{n\to m}}$$

and $u_{n,m}$ is the proposal values that "complete" the parameter values from dimension $n$ when transformed to dimension $m$.

To ensure the detailed-balance property, reversibility of the Markov chain must be maintained. This means that at every iteration, a jump from dimension $m$ to $n$ must be followed by a reverse jump from dimension $n$ to $m$. In between, there must be a step to update the parameter values within dimensions.

According to (Sillanpää and Arjas 1998), the steps in each iteration in the QTL analysis are:

1. QTL birth, *i.e.,* introducing one QTL into the model
2. Moving one QTL position at random
3. Sampling QTL effects
4. QTL death, *i.e.,* deleting one QTL at random from the model

The first and last steps are implemented as RJMCMC acceptance tests. The others are implemented in MH algorithms.

In each of these steps, the ratio of the posterior probabilities reduces to the ratio of likelihoods of the two states. For both QTL birth and death, the Jacobian term of the RJMCMC is 1 since the locations of existing QTLs do not determine the location of a proposed new QTL, nor does the deletion of one QTL influence the positions of the remaining QTLs. The ratios $q(n|m)/q(m|n)$ in QTL birth and death steps are $\lambda/(N_q^{(t-1)} + 1)^2$ and $(N_q^{(t-1)})^2/\lambda$, where $N_q^{(t-1)}$ is the number of QTLs in the previous iteration and $\lambda$ is the prior

belief of the number of QTLs. Although (Sillanpää and Arjas 1998) gave a formula to estimate $\lambda$, it can be estimated by the number of markers given by a cofactor-marker selection process. The ratio of the proposal distribution can be used to restrict the search space, as in

$$\frac{J_{m\to n}(u_{m,n}|m,\boldsymbol{\theta}_m)]}{J_{n\to m}(u_{n,m}|n,\boldsymbol{\theta}_n)]} = \begin{cases} 1, & \text{if } 0 \le N_q^{(t)} \le N_{qmax} \\ 0, & \text{otherwise} \end{cases}$$

where $N_{qmax}$ is the maximum number of QTLs allowed in the model. When the position or the effect of a QTL are being updated, the ratio of the proposal distributions is assumed to be 1.

The first $b$ iterations are discarded since the Markov chain has not reached convergence. After the first $b$ iterations, only samples at every $t$ iterations are stored to avoid autocorrelation in samples between adjacent iterations. If $n$ samples are required, the steps are run $nt + b$ times. In MCMC term, $b$ represents *burn-in*, and $t$ *thinning* iterations.

There is no consensus on how $b$ and $t$ should be set. Although rules have been proposed, such as by Raftery and Lewis (1995), it is still a contentious subject. Sillanpää and Arjas (1998) suggest 500,000 to 1,500,000 iterations with no burn-in or thinning iterations. The default setting in R/qtlbim (Yi *et al.* 2005) is $b = 600$, $t = 20$, and $n = 3000$. However, in most cases such defaults will be unsuitable for the data. The only way to determine the setting is to examine the plot of the number of QTLs, general mean, and the residual variance across iterations.

## 2.7 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 cofactors.

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.

# Chapter 3

# Shrinkage interval mapping


## Abstract

The conventional solution to the QTL-mapping model-selection problem, in which only a few QTLs are selected from all QTLs, obtained from stepwise or forward variable-selection methods has been shown to perform poorly due to bias introduced by favoring QTLs that are associated with the largest statististics. A proposed remedy to this problem is penalized regression, in particular the penalized maximum-likelihood method (PMLE). However, this method tends to overpenalize, and thereby may fail to detect, QTLs with smaller effects. As an attempt to overcome this defect, I develop a two-stage hybrid method between MIM and partially penalized regression that can be considered as a generalization of PMLE. A multiple-trait extension is also developed. Simulated experiments showed that it may obtain a more precise QTL-location estimate, but may overestimate the QTL effects. This method has a marginal advantage over PMLE in detecting QTLs with smaller effects. Both PMLE and this method (shrinkIM) are shown to be superior to multiple-interval mapping (MIM).


## 3.1 Introduction

Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is a $n \times 1$ trait vector, $\mathbf{X}$ a $n \times m$ marker matrix, $\boldsymbol{\beta}$ a $m \times 1$ vector of regression coefficients, and $\epsilon$ a $n \times 1$ random error vector with

$\epsilon \sim N(0, \mathbf{I}_n\sigma^2)$. For an oversaturated model, where $m > n$, ordinary least-squares estimates of $\boldsymbol{\beta}$ cannot be calculated as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ because matrix $\mathbf{X}'\mathbf{X}$ is singular.

Composite interval mapping (CIM) was proposed to overcome this situation by stepwise or forward variable-selection methods. However, it has been shown (Hoerl *et al.* 1986) that such methods perform poorly due to the bias introduced by favoring variables (or in this case QTL) that are associated with the largest statistics.

Another solution, ridge regression (Hoerl 1962), proposed to overcome this problem is the imposition of penalties on the regression coefficients. Let $\tau$ be a penalty parameter. The restricted least-squares estimate is $(\mathbf{X}'\mathbf{X} + \tau\mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{Y}$ under the quadratic constraint $\sum_j \beta_j^2 < \tau$ on $\boldsymbol{\beta}$, for $j = 1, \ldots, m$. Ridge regression has been proposed (Boer *et al.* 2002) for QTL epistasis analysis, with varying penalties on regression coefficients.

Because the inversion of matrix $\mathbf{X}'\mathbf{X} + \tau\mathbf{I}_n$ required by ridge regression becomes time-consuming as $m$ grows, Xu (2003) developed a Bayesian shrinkage regression method for simultaneously estimating the genetic effect associated with the markers along the whole genome map. Each marker effect is allowed to have its own variance parameters so that the variance can be estimated from the data. This method was extended (Wang *et al.* 2005) to allow localizing a QTL within an interval, using Metropolis–Hastings sampling since the QTL location parameter does not have an explicit posterior distribution.

To eliminate the need of intensive computation imposed by the Bayesian method, the penalized maximum-likelihood estimation (PMLE) method (Zhang and Xu 2005) was developed. It imposes a prior normal $N(\mu_j, \sigma_j^2)$ penalty on each QTL effect $j$, allowing the penalty to vary across the $\beta_j$. An EM-based algorithm is used to estimate regression coefficients and other parameters. PMLE is similar in spirit to the multiple-marker Bayesian shrinkage method in that both shrink small marker or QTL effects to zero. PMLE was shown (Zhang and Xu 2005) to be comparable to the shrinkage method in terms of performance. The initial PMLE method could localize a QTL only to a marker and not between markers. An extension was developed (Zhang 2006) to accommodate QTLs within intervals.

While both shrinkage and PMLE methods offer much power for QTL detection, the QTL effects can sometimes be underestimated (Zhang and Xu 2005). Although this is not serious if the effect is large, it can be highly misleading for small effects.

To overcome the limitation of PMLE, an unpublished method called shrinkage interval mapping (shrinkIM) (Guo *et al.* 2007) was proposed. It used PMLE as QTL selector and used unpenalized QTL effect estimates to find other QTLs with smaller effects. This method exploited partially penalized regression to leave QTL effect estimates unpenalized while penalizing spurious effects.

Here, I propose an improvement to shrinkIM that combines PMLE with multiple interval mapping (MIM) (Kao *et al.* 1999). This method is a multiple-pass method. In the first pass, PMLE is used to detect QTLs with higher effects. In the second pass, a hybrid between MIM and partially penalized regression is used to fit without penalty the QTLs found in the first pass while simultaneously searching for additional QTLs with smaller effects. This simultaneous unpenalized QTL fitting is analogous to that of MIM and is intended to improve precision and power. Further passes repeat the second until no more QTLs are found.

## 3.2 Methods of shrinkage interval mapping

Consider $m$ QTL, $Q_1, \ldots, Q_m$, located at positions $p_1, \ldots, p_m$ of $m$ intervals across the genome that have been not identified by the PMLE method in the previous iteration. Assume an $F_2$ population of $n$ individuals. There are $3^m$ possible different QTL genotypes in the population. These QTL genotypes determine a quantitative trait $y$. Assuming no interaction, the statistical model can be expressed as

$$y_i = \mu + \mathbf{x}_i\boldsymbol{\beta} + \sum_{j=1}^{m}(\alpha_j z_{ij} + \delta_j w_{ij}) + \epsilon_i, \quad i = 1 \ldots n, \tag{3.1}$$

or in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\mu$ is the mean, $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are the genotypes and unpenalized effects of the QTL found in previous iterations, $z_{ij}$ is coded as $1, 0$, and $-1$ for current QTL genotypes $Q_jQ_j$, $Q_jq_j$, and $q_jq_j$, respectively, $w_{ij}$ is coded as $\frac{1}{2}$ for heterozygotes and $-\frac{1}{2}$ otherwise, $\alpha_j$ and $\delta_j$ are the additive and dominance effects of QTL $j$, and $\epsilon_i$ is the environment deviation, assumed to follow $N(0, \sigma^2)$.

The encoding of $z_{ij}$ and $w_{ij}$ follows Cockerham's model (Kao and Zeng 2002), which ensures orthogonality in modeling the genetic parameters. It means that we can treat the $\alpha$s and $\delta$s as independent variables.

The putative QTL genotypes are not observed because the QTL lie within marker intervals. Given observed flanking marker genotypes, the conditional distributions of QTL genotypes can be inferred with a Markov chain method (Jiang and Zeng 1997), assuming no crossover interference.

The model is a multiple-QTL model and its likelihood is that of a finite normal mixture. Let $\mathbf{q}$ be the vector of QTL genotypes, $\boldsymbol{\lambda}$ the vector of the positions of the QTL, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ the vectors of additive and dominance effects of all $m$ QTLs, $\boldsymbol{\beta}$ the effects of the confirmed QTLs, and $\sigma^2$ be the variance. Let $\boldsymbol{\kappa} = (\mathbf{q}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2)$. Assuming that the putative QTL genotypes, given the flanking marker genotypes, are independent, that no penalty is imposed, and that noninformative priors are chosen, the posterior probability of the parameters is

$$\begin{aligned} p(\boldsymbol{\kappa}|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}) \\ &= p(\mathbf{y}|\boldsymbol{\kappa})p(\mathbf{q}|\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2) \\ &= p(\mathbf{y}|\boldsymbol{\kappa})p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)\prod_{j=1}^{m} p(q_j|\lambda_j)p(\alpha_j)p(\delta_j) \qquad (3.3a) \\ &\propto \frac{1}{\sigma^2}p(\mathbf{y}|\boldsymbol{\kappa}) \qquad (3.3b) \end{aligned}$$

We impose a penalty on each QTL with the corresponding hierarchical model

$$\alpha_j|\mu_{\alpha_j}, \sigma^2_{\alpha_j} \sim N(\mu_{\alpha_j}, \sigma^2_{\alpha_j}) \qquad (3.4a)$$

$$\delta_j|\mu_{\delta_j},\sigma^2_{\delta_j} \sim N(\mu_{\delta_j},\sigma^2_{\delta_j}) \tag{3.4b}$$

$$\mu_{\alpha_j}|\sigma^2_{\alpha_j},\eta \sim N(0,\sigma^2_{\alpha_j}/\eta) \tag{3.4c}$$

$$\mu_{\delta_j}|\sigma^2_{\delta_j},\eta \sim N(0,\sigma^2_{\delta_j}/\eta) \tag{3.4d}$$

$$p(\sigma^2_{\alpha_j}) \propto \sigma^{-2}_{\alpha_j} \tag{3.4e}$$

$$p(\sigma^2_{\delta_j}) \propto \sigma^{-2}_{\delta_j} \tag{3.4f}$$

Including the penalty, we may rewrite (3.3a) as

$$\begin{aligned} p(\boldsymbol{\kappa}|\mathbf{y}) &= p(\mathbf{y}|\boldsymbol{\kappa})p(\boldsymbol{\beta},\sigma^2)\prod_{j=1}^{m} p(q_j|\boldsymbol{\lambda})p(\alpha_j|\mu_{\alpha_j},\sigma^2_{\alpha_j})p(\mu_{\alpha_j},\sigma^2_{\alpha_j})p(\delta_j|\mu_{\delta_j},\sigma^2_{\delta_j})p(\mu_{\delta_j},\sigma^2_{\delta_j}) \\ &= p(\mathbf{y}|\boldsymbol{\kappa})p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\ &\times \prod_{j=1}^{m} p(q_j|\boldsymbol{\lambda})p(\alpha_j|\mu_{\alpha_j},\sigma^2_{\alpha_j})p(\delta_j|\mu_{\delta_j},\sigma^2_{\delta_j})p(\mu_{\alpha_j}|0,\sigma^2_{\alpha_j}/\eta)p(\mu_{\delta_j}|0,\sigma^2_{\delta_j}/\eta) \end{aligned} \tag{3.5}$$

where $\mu_{\alpha_j}$, $\mu_{\delta_j}$, $\sigma^2_{\alpha_j}$, and $\sigma^2_{\delta_j}$ are the prior mean and variance values of the additive and dominance effects, respectively, and $\eta > 0$ is a penalty parameter.

The likelihood function of the model is

$$M(\boldsymbol{\kappa}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\kappa}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\kappa})$$

Let $\boldsymbol{\xi} = (\mu_{\alpha_1},\mu_{\delta_1},\sigma^2_{\alpha_1},\sigma^2_{\delta_1},\ldots,\mu_{\alpha_m},\mu_{\delta_m},\sigma^2_{\alpha_m},\sigma^2_{\delta_m})$ be the penalty hyperparameters. The prior density used as the penalty is

$$P(\boldsymbol{\kappa},\boldsymbol{\xi}|\mathbf{y}) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)\prod_{j=1}^{m} p(q_j|\boldsymbol{\lambda})p(\alpha_j|\mu_{\alpha_j},\sigma^2_{\alpha_j})p(\delta_j|\mu_{\delta_j},\sigma^2_{\delta_j})p(\mu_{\alpha_j}|0,\sigma^2_{\alpha_j}/\eta)p(\mu_{\delta_j}|0,\sigma^2_{\delta_j}/\eta)$$

Let $\boldsymbol{\theta} = (\boldsymbol{\kappa},\boldsymbol{\xi})$ The partially penalized likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = M(\boldsymbol{\kappa}|\mathbf{y})P(\boldsymbol{\kappa},\boldsymbol{\xi}|\mathbf{y})$$

Observe that the partially penalized likelihood function estimates the priors with the likelihood at the same time. Thus, partially penalized maximum-likelihood estimation

(PPMLE) can be thought of as a Bayesian method that imposes a prior distribution on the QTL effects.

The likelihood function with penalty based on (3.5) is

$$
\begin{aligned}
L(\boldsymbol{\theta}|\mathbf{y}) &= p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)\left[\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta},\mathbf{q})p(\mathbf{q}|\boldsymbol{\lambda})\right]\left[\prod_{j=1}^{m} p(\alpha_j|\mu_{\alpha_j},\sigma^2_{\alpha_j})p(\delta_j|\mu_{\delta_j},\sigma^2_{\delta_j})\right] \\
&\times\left[\prod_{j=1}^{m} p(\mu_{\alpha_j}|0,\sigma^2_{\alpha_j}/\eta)p(\mu_{\delta_j}|0,\sigma^2_{\delta_j}/\eta)\right] \qquad (3.6)\\
&\propto \frac{1}{\sigma^2}\left[\prod_{i=1}^{n}\sum_{k=1}^{3^m} p_{ik}\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ik}}{\sigma}\right)\right]\left[\prod_{j=1}^{m}\frac{1}{\sigma_{\alpha_j}}\phi\left(\frac{\alpha_j-\mu_{\alpha_j}}{\sigma_{\alpha_j}}\right)\frac{1}{\sigma_{\delta_j}}\phi\left(\frac{\delta_j-\mu_{\delta_j}}{\sigma_{\delta_j}}\right)\right. \\
&\left.\times\frac{\sqrt{\eta}}{\sigma_{\alpha_j}}\phi\left(\frac{\mu_{\alpha_j}}{\sigma_{\alpha_j}/\sqrt{\eta}}\right)\frac{\sqrt{\eta}}{\sigma_{\delta_j}}\phi\left(\frac{\mu_{\delta_j}}{\sigma_{\delta_j}/\sqrt{\eta}}\right)\right] \qquad (3.7)
\end{aligned}
$$

where $\phi(\cdot)$ is the standard normal pdf, and $\mu_{ik}$ is the effect of the $k^{\text{th}}$ combination of $m$ QTL genotypes. Let $\mathbf{1} = (1,1,1)'$ and $\boldsymbol{\xi}_j = (\alpha_j - \delta_j/2, \delta_j/2, -\alpha_j - \delta_j/2)'$. Let $\boldsymbol{\Xi}_j = \underbrace{\mathbf{1}\otimes\ldots\otimes\mathbf{1}\otimes\boldsymbol{\xi}_j}_{j \text{ times}}\otimes\overbrace{\mathbf{1}\ldots\otimes\mathbf{1}}^{m \text{ times}}$, where $\otimes$ is the Kronecker product. Let $\mathbf{K} = \sum_{j=1}^{m}\boldsymbol{\Xi}_j$. Note that $\boldsymbol{\Xi}_j$ and $\mathbf{K}$ are vectors of $3^m$ elements. Let $\mathbf{K}_k$ be the $k^{\text{th}}$ element of $\mathbf{K}$. Then $\mu_{ik} = \mathbf{K}_k + \mathbf{x}_i\boldsymbol{\beta}$.

We solve this likelihood equation using an ECM (Meng and Rubin 1993) algorithm, an extension of the EM algorithm (Dempster *et al.* 1977). Let $y_{\text{mis}}$ and $y_{\text{obs}}$ stand for missing data and observed data, respectively. We treat the unobserved QTL genotypes as missing data. Let $L(\boldsymbol{\theta}|y_{\text{com}})$ be the likelihood for complete data and $f(y)$ be the joint pdf of $y$. The formulation of the E step involves calculation of the conditional expected complete-data log-likelihood with respect to the conditional distribution of $y_{\text{mis}}$ given $y_{\text{obs}}$. The current estimated parameter value $\boldsymbol{\theta}^{(t)}$ is

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \int \log L(\boldsymbol{\theta}|\mathbf{y}_{\text{com}}) f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)})\, d\mathbf{y}_{\text{mis}} \\
&= \int\left\{\log\left[\prod_{i=1}^{n} p_i\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right] + \sum_{j=1}^{m}\left[-\frac{1}{2}\log\sigma^2_{\alpha_j} + \log\phi\left(\frac{\alpha_j-\mu_{\alpha_j}}{\sigma_{\alpha_j}}\right)\right.\right. \\
&\left.\left. -\frac{1}{2}\log\sigma^2_{\delta_j} + \log\phi\left(\frac{\delta_j-\mu_{\delta_j}}{\sigma_{\delta_j}}\right) - \frac{1}{2}\log\frac{\sigma^2_{\alpha_j}}{\eta} + \log\phi\left(\frac{\mu_{\alpha_j}}{\sigma_{\alpha_j}/\sqrt{\eta}}\right)\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{2}\log\frac{\sigma^2_{\delta_j}}{\eta}+\log\phi\left(\frac{\mu_{\delta_j}}{\sigma_{\delta_j}/\sqrt{\eta}}\right)\Bigg]\Bigg\}\times f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)})\,d\mathbf{y}_{\text{mis}}\\
&=\int\left\{\log\left[\prod_{i=1}^{n}p_i\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]+\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta)\right\}\\
&\times f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)})\,d\mathbf{y}_{\text{mis}}\\
&=\int\left[\log\prod_{i=1}^{n}p_i\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)})\,d\mathbf{y}_{\text{mis}}\\
&+\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta),
\end{aligned}
$$

where $\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta)$ is shorthand for the parameters.

By Fubini's theorem governing changes in order of integration,

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \left[\sum_{i=1}^{n}\sum_{j=1}^{3^m}\log\left[p_{ij}\ \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ij}}{\sigma}\right)\right]\times f(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)})\right]\\
&+\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta)\\
&=\left[\sum_{i=1}^{n}\sum_{j=1}^{3^m}\log\left[p_{ij}\ \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ij}}{\sigma}\right)\right]\times\frac{p_{ij}\ \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ij}}{\sigma}\right)}{\sum_{k=1}^{3^m}p_{ik}\ \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ik}}{\sigma}\right)}\right]\\
&+\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta)\\
&=\left[\sum_{i=1}^{n}\sum_{j=1}^{3^m}\log\left[p_{ij}\ \frac{1}{\sigma}\phi\left(\frac{y_i-\mu_{ij}}{\sigma}\right)\right]\pi_{ij}\right]+\psi(\boldsymbol{\mu}_\alpha,\boldsymbol{\sigma}^2_\alpha,\boldsymbol{\mu}_\delta,\boldsymbol{\sigma}^2_\delta,\eta) \qquad (3.8)
\end{aligned}
$$

where $\pi_{ij}$ is the posterior probability of the QTL genotype. Thus, we can view the E step as updating $\pi_{ij}$.

Let $\alpha_j$ and $\delta_j$ be the additive and dominance effects of the $j^{\text{th}}$ QTL. The M step is

$$
\alpha_j^{(t+1)}=\frac{\left[\sum_{i=1}^{n}(y_i-x_i\beta^{(t)})(\pi_{i1}^{(t)}-\pi_{i3}^{(t)})+\frac{1}{2}(\pi_{i3}^{(t)}-\pi_{i1}^{(t)})\delta_j^{(t)}\right]+\frac{\mu_{\alpha_j}^{(t)}\sigma^{2(t)}}{\sigma_{\alpha_j}^{2(t)}}}{\left[\sum_{i=1}^{n}(\pi_{i1}^{(t)}+\pi_{i3}^{(t)})\right]+\frac{\sigma^{2(t)}}{\sigma_{\alpha_j}^{2(t)}}} \qquad (3.9a)
$$

$$\delta_j^{(t+1)} = \frac{\left[\sum_{i=1}^{n} \frac{1}{2}(y_i - x_i\beta^{(t)})(-\pi_{i1}^{(t)} + \pi_{i2}^{(t)} - \pi_{i3}^{(t)}) - \frac{1}{2}(\pi_{i3}^{(t)} - \pi_{i1}^{(t)})\alpha_j^{(t)}\right] + \frac{\mu_{\delta_j}^{(t)}\sigma^{2(t)}}{\sigma_{\delta_j}^{2(t)}}}{\left[\sum_{i=1}^{n}(\pi_{i1}^{(t)} + \pi_{i2}^{(t)} + \pi_{i3}^{(t)})\right] + \frac{\sigma^{2(t)}}{\sigma_{\delta_j}^{2(t)}}} \tag{3.9b}$$

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{y} - \left(\sum_{j=1}^{m}\mathbf{Z}_j\alpha_j^{(t+1)} + \mathbf{W}_j\delta_j^{(t+1)}\right)\right] \tag{3.9c}$$

$$\boldsymbol{S}^{(t+1)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \left(\sum_{j=1}^{m}\mathbf{Z}_j\alpha_j^{(t+1)} + \mathbf{W}_j\delta_j^{(t+1)}\right)$$

$$\sigma^{2(t+1)} = \frac{1}{n}\boldsymbol{S}^{(t+1)'}\boldsymbol{S}^{(t+1)} \tag{3.9d}$$

$$\mu_{\alpha_j}^{(t+1)} = \frac{\alpha_j^{(t+1)}}{\eta + 1} \tag{3.9e}$$

$$\sigma_{\alpha_j}^{2(t+1)} = \frac{1}{2}[(\alpha_j^{(t+1)} - \mu_{\alpha_j}^{(t)})^2 + \eta\mu_{\alpha_j}^{2(t)}] \tag{3.9f}$$

$$\mu_{\delta_j}^{(t+1)} = \frac{\delta_j^{(t+1)}}{\eta + 1} \tag{3.9g}$$

$$\sigma_{\delta_j}^{2(t+1)} = \frac{1}{2}[(\delta_j^{(t+1)} - \mu_{\delta_j}^{(t)})^2 + \eta\mu_{\delta_j}^{2(t)}] \tag{3.9h}$$

Initially, we set $\alpha_j = \delta_j = \mu_{\alpha_j} = \mu_{\delta_j} = 0$ and $\sigma_{\alpha_j}^2 = \sigma_{\delta_j}^2 = s^2$ for all $j \in \{1, 2, \ldots, m\}$, where $s^2$ is the variance of $\mathbf{y}$.

Note that $\frac{\mu_k^{(t)}\sigma^{2(t)}}{\sigma_k^{2(t)}}$ and $\frac{\sigma^{2(t)}}{\sigma_k^{2(t)}}$ are adjustment terms for the $k^{\text{th}}$ effect. Intuitively, if the variance of that effect, $\sigma_k^2$, is large, these terms do not much alter the effect estimate. If it is small, the effect will be shrunk to zero. Since $\mu_k$ is defined as a portion of the $k^{\text{th}}$ effect on the previous iteration, $\sigma_k^2$ is effectively the average of the square of the difference of effect estimates between the previous and current iteration and the square of the effect of the previous iteration. The $\sigma_k^2$ estimate will be small if the difference of effect estimates between iterations is small. This situation occurs when the initial value of $\mu_k$ is close to the estimate. Thus, if the initial value of $\mu_k$ is zero, all effects close to zero will be shrunk to zero. The penalty parameter $\eta > 0$ controls the sensitivity of the shrinking process. The

closer $\eta$ is to zero, the more sensitive the process is to detect QTLs with small effects, at the expense of slower convergence.

This can be considered as a generalization of the PMLE method. The major differences lie in the $\mathbf{X}$ matrix, holding the estimated QTL genotype found in the previous round, and $\boldsymbol{\beta}$ vector, holding unpenalized QTL effect estimates. In the first round, $\mathbf{X} = \mathbf{0}$, reducing PPMLE to PMLE. In subsequent rounds, QTLs found in the previous rounds are fitted simultaneously and in unpenalized fashion while additional QTLs with smaller effects are sought.

## 3.3 Hypothesis test for univariate case

The testing of hypotheses is not possible in an oversaturated model because of overparameterization. I follow a two-stage selection process proposed by (Zhang and Xu 2005). In the first stage, we select all QTLs with either $|\alpha_j|/\hat{\sigma} > 10^{-6}$ or $|\delta_j|/\hat{\sigma} > 10^{-6}$. In the second stage, we compute the likelihood ratio statistic for each QTL with $\mathrm{LRT}_j = -2[L(\boldsymbol{\theta}_{-j}) - L(\boldsymbol{\theta})]$, where $\boldsymbol{\theta}_{-j}$ is all parameters with $\alpha_j$ and $\delta_j$ set to 0. Alternatively, the LOD score statistic may be used. It is calculated as LRT/(2 ln 10).

## 3.4 Generalization for epistatic effect

Let # denote the Hadamard product (componentwise product of two vectors), $\mathbf{\Pi} = \{\pi_{ji}\}_{n\times 3}$, $\mathbf{d}_1 = (1, 0, -1)'$, $\mathbf{d}_2 = (-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2})'$, and $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2)$. Let $\mathbf{e}^{(t)} = (\alpha_j^{(t)}, \delta_j^{(t)})'$, for $j \in \{1, 2, \ldots, m\}$. Equations (3.9a–h) simplify to

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

where

$$\mathbf{r}^{(t)} = \begin{bmatrix} \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_1+\frac{\mu_{\alpha_j}^{(t)}\sigma^{2(t)}}{\sigma^{2(t)}_{\alpha_j}}}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1)+\frac{\sigma^{2(t)}}{\sigma^{2(t)}_{\alpha_j}}} \\ \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_2+\frac{\mu_{\alpha_j}^{(t)}\sigma^{2(t)}}{\sigma^{2(t)}_{\delta_j}}}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2)+\frac{\sigma^{2(t)}}{\sigma^{2(t)}_{\delta_j}}} \end{bmatrix} \quad \text{and} \quad \mathbf{M}^{(t)} = \begin{bmatrix} 0 & \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_2)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1\#\mathbf{d}_1)+\frac{\sigma^{2(t)}}{\sigma^{2(t)}_{\alpha_j}}} \\ \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_1)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2\#\mathbf{d}_2)+\frac{\sigma^{2(t)}}{\sigma^{2(t)}_{\delta_j}}} & 0 \end{bmatrix}$$

and

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\Pi}^{(t)}\mathbf{D}\mathbf{e}^{(t+1)}) \tag{3.10a}$$

$$\sigma^{2(t+1)} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\Pi}^{(t)}\mathbf{D}\mathbf{e}^{(t+1)})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\Pi}^{(t)}\mathbf{D}\mathbf{e}^{(t+1)}) \tag{3.10b}$$

$$\boldsymbol{\mu}_{\mathbf{e}}^{(t+1)} = \frac{\mathbf{e}^{(t+1)}}{\eta+1} \tag{3.10c}$$

$$\boldsymbol{\sigma}_{\mathbf{e}}^{2(t+1)} = \frac{1}{2}\text{Diag}[(\mathbf{e}^{(t+1)} - \boldsymbol{\mu}_{\mathbf{e}}^{(t)})'(\mathbf{e}^{(t+1)} - \boldsymbol{\mu}_{\mathbf{e}}^{(t)}) + \eta\boldsymbol{\mu}_{\mathbf{e}}^{(t)'}\boldsymbol{\mu}_{\mathbf{e}}^{(t)}] \tag{3.10d}$$

where $\mathbf{e}$, $\boldsymbol{\mu}_{\mathbf{e}}$, and $\boldsymbol{\sigma}^2_{\mathbf{e}}$ are $2m \times 1$ vectors.

With this notation, it is straightforward to extend the model for epistatic interactions. Still assuming $F_2$, there are 8 effects per QTL with 9 possible genotypes. The new genetic design matrix $\mathbf{D}^*$ is a $9\times 8m$ matrix to reflect the computation of the effects. Let $\mathbf{j} = (1,1,1)'$. Write

$$\mathbf{D}^* = [\mathbf{D}\otimes\mathbf{j}|\mathbf{j}\otimes\mathbf{D}|\mathbf{D}\otimes\mathbf{D}]$$

So, for $F_2$, $\mathbf{D}^*$ for one QTL is

$$\mathbf{D}^* = \begin{bmatrix} 1 & -\frac{1}{2} & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ 1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{4} \\ 1 & -\frac{1}{2} & -1 & -\frac{1}{2} & -1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & -1 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & -\frac{1}{4} \\ -1 & -\frac{1}{2} & 1 & -\frac{1}{2} & -1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ -1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & -\frac{1}{4} \\ -1 & -\frac{1}{2} & -1 & -\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

Let $\mu_i$ and $\sigma_i^2$ be the $i^{\text{th}}$ elements of $\mu_{\mathbf{e}}$ and $\sigma^2_{\mathbf{e}}$, respectively. Let $\mathbf{d}_i^*$ be the $i^{\text{th}}$ column of $\mathbf{D}^*$. Thus, $\mathbf{r}^{(t)} = [r_i]$ and $\mathbf{M}^{(t)} = [m_{ij}]$ are an $8m\times 1$ vector and an $8m\times 8m$ matrix, respectively,

where

$$r_i = \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_i^* + \frac{\mu_i^{(t)} \sigma^{2(t)}}{\sigma_i^{2(t)}}}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_i^* \# \mathbf{d}_i^*) + \frac{\sigma^{2(t)}}{\sigma_i^{2(t)}}}$$

and

$$m_{ij} = \frac{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_i^* \# \mathbf{d}_j^*)}{\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_i^* \# \mathbf{d}_i^*) + \frac{\sigma^{2(t)}}{\sigma_i^{2(t)}}}, \qquad \text{if } i \neq j, \ 0 \text{ otherwise.}$$

$\boldsymbol{\Pi}$ is now an $n \times 9$ matrix, where each row expresses the probability of genotypes $AABB$, $AABb$, $AAbb$, $AaBB$, $AaBb$, $Aabb$, $aaBB$, $aaBb$, and $aabb$ of each individual.

It is also possible to obtain formulas for higher-order epistatic interaction. Each column in the genetic design matrix will be a permutation of Kronecker products. For example, for three-way epistatic interaction, the genetic design matrix is

$$\mathbf{D}^* = [\mathbf{D} \otimes \mathbf{j} \otimes \mathbf{j} | \mathbf{j} \otimes \mathbf{D} \otimes \mathbf{j} | \mathbf{j} \otimes \mathbf{j} \otimes \mathbf{D} | \mathbf{D} \otimes \mathbf{D} \otimes \mathbf{j} | \mathbf{D} \otimes \mathbf{j} \otimes \mathbf{D} | \mathbf{j} \otimes \mathbf{D} \otimes \mathbf{D}] | [\mathbf{D} \otimes \mathbf{D} \otimes \mathbf{D}]$$

Vector $\mathbf{r}$, matrix $\mathbf{M}$, and matrix $\boldsymbol{\Pi}$ are modified similarly.

## 3.5 Multivariate extension

In multivariate extension, the scalar $y_i$ in the univariate model becomes $\mathbf{y}_i$, a vector of $t$ elements. Thus, the model becomes

$$\mathbf{y}_i = \boldsymbol{\mu} + (\mathbf{x}_i \mathbf{B})' + \sum_{j=1}^{m} (\boldsymbol{\alpha}_j z_{ij} + \boldsymbol{\delta}_j w_{ij}) + \boldsymbol{\epsilon}_i, \quad i = 1 \ldots n, \tag{3.11}$$

where $\boldsymbol{\mu}, \boldsymbol{\alpha}_j$, and $\boldsymbol{\delta}_j$ are vectors of $t$ elements denoting the mean, additive, and dominance effects, $(\mathbf{x}_i \mathbf{B})'$ are the $t \times 1$ unpenalized estimates of QTLs found in the previous iterations, and $\boldsymbol{\epsilon}_i$ is the environment deviation, assumed to follow $N(\mathbf{0}, \boldsymbol{\Sigma})$.

Let $\boldsymbol{\phi}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the multivariate normal pdf with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ and $\mathbf{I}_{2e}$ a $2e \times 2e$ identity matrix, where $e$ is the number of QTLs found in previous iterations. The likelihood function is

$$L(\boldsymbol{\theta}|y) \propto p(\boldsymbol{B}|\boldsymbol{\Sigma} \otimes \mathbf{I}_{2e}) p(\boldsymbol{\Sigma}) \left[\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}, \mathbf{q}) p(\mathbf{q}|\boldsymbol{\lambda})\right] \left[\prod_{j=1}^{m} p(\boldsymbol{\alpha}_j|\boldsymbol{\mu}_{\alpha_j}, \mathbf{V}_{\alpha_j}) p(\boldsymbol{\delta}_j|\boldsymbol{\mu}_{\delta_j}, \mathbf{V}_{\delta_j})\right]$$

$$= p(\boldsymbol{B}|\boldsymbol{\Sigma} \otimes \mathbf{I}_{2e}) p(\boldsymbol{\Sigma}) \left[ \prod_{i=1}^{n} \sum_{k=1}^{3^m} p_{ik} \boldsymbol{\phi}(\mathbf{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}) \right] \left[ \prod_{j=1}^{m} \boldsymbol{\phi}(\boldsymbol{\alpha}_j; \boldsymbol{\mu}_{\alpha_j}, \mathbf{V}_{\alpha_j}) \boldsymbol{\phi}(\boldsymbol{\delta}_j; \boldsymbol{\mu}_{\delta_j}, \mathbf{V}_{\delta_j}) \right.$$
$$\left. \times \boldsymbol{\phi}(\boldsymbol{\mu}_{\alpha_j}; \mathbf{0}, \mathbf{V}_{\alpha_j}/\eta) \boldsymbol{\phi}(\boldsymbol{\mu}_{\delta_j}; \mathbf{0}, \mathbf{V}_{\delta_j}/\eta) \right] \tag{3.12}$$

The derivation is similar to the univariate case. Let $e$ be the number of effects per QTL. Using the same notation as the one in epistatic extension, the result is

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

where $\mathbf{R}^{(t)}$ is an $emt \times t$ matrix

$$\mathbf{R}^{(t)} = \begin{bmatrix} \left\{ \left( \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)} \right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_1 + \boldsymbol{\mu}_{\alpha_1}^{(t)} \mathbf{V}_{\alpha_1}^{-1(t)} \mathbf{V}^{(t)} \right\} \times \left\{ \left[ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_1 \# \mathbf{d}_1) \right] \mathbf{I}_t + \mathbf{V}_{\alpha_1}^{-1(t)} \mathbf{V}^{(t)} \right\}^{-1} \\ \left\{ \left( \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)} \right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_2 + \boldsymbol{\mu}_{\delta_1}^{(t)} \mathbf{V}_{\delta_1}^{-1(t)} \mathbf{V}^{(t)} \right\} \times \left\{ \left[ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_2 \# \mathbf{d}_2) \right] \mathbf{I}_t + \mathbf{V}_{\delta_1}^{-1(t)} \mathbf{V}^{(t)} \right\}^{-1} \\ \vdots \\ \left\{ \left( \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)} \right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_1 + \boldsymbol{\mu}_{\alpha_m}^{(t)} \mathbf{V}_{\alpha_m}^{-1(t)} \mathbf{V}^{(t)} \right\} \times \left\{ \left[ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_1 \# \mathbf{d}_1) \right] \mathbf{I}_t + \mathbf{V}_{\alpha_m}^{-1(t)} \mathbf{V}^{(t)} \right\}^{-1} \\ \left\{ \left( \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)} \right)' \boldsymbol{\Pi}^{(t)} \mathbf{d}_2 + \boldsymbol{\mu}_{\delta_m}^{(t)} \mathbf{V}_{\delta_m}^{-1(t)} \mathbf{V}^{(t)} \right\} \times \left\{ \left[ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_2 \# \mathbf{d}_2) \right] \mathbf{I}_t + \mathbf{V}_{\delta_m}^{-1(t)} \mathbf{V}^{(t)} \right\}^{-1} \end{bmatrix}$$

and $\mathbf{M}^{(t)}$ is an $emt \times emt$ block matrix. Let $\mathbf{M}_{ij}^{(t)}$ be a $et \times et$ matrix from $m$ equal partitions of $\mathbf{M}^{(t)}$.

$$\mathbf{M}_{ij}^{(t)} = \left\{ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_i^* \# \mathbf{d}_j^*) \right\} \times \left\{ \left[ \mathbf{1}' \boldsymbol{\Pi}^{(t)} (\mathbf{d}_i \# \mathbf{d}_i) \right] \mathbf{I}_t + \mathbf{V}_i^{-1(t)} \mathbf{V}^{(t)} \right\}^{-1} \text{ if } i \neq j, \ 0 \text{ otherwise.}$$

and

$$\mathbf{B}^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\mathbf{Y} - \boldsymbol{\Pi}^{(t)} \mathbf{D} \mathbf{E}^{(t+1)}) \tag{3.13a}$$

$$\mathbf{V}^{(t+1)} = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\mathbf{B} - \boldsymbol{\Pi}^{(t)} \mathbf{D} \mathbf{E}^{(t+1)})' (\mathbf{Y} - \mathbf{X}\mathbf{B} - \boldsymbol{\Pi}^{(t)} \mathbf{D} \mathbf{E}^{(t+1)}) \tag{3.13b}$$

$$\boldsymbol{\mu}_{\mathbf{E}}^{(t+1)} = \frac{\mathbf{E}^{(t+1)}}{\eta + 1} \tag{3.13c}$$

$$\mathbf{V}_i^{(t+1)} = \frac{1}{2} \text{Diag}[(\mathbf{E}^{(t+1)} - \boldsymbol{\mu}_{\mathbf{E}}^{(t)})' (\mathbf{E}^{(t+1)} - \boldsymbol{\mu}_{\mathbf{E}}^{(t)}) + \eta \boldsymbol{\mu}_{\mathbf{E}}^{(t)'} \boldsymbol{\mu}_{\mathbf{E}}^{(t)}], \tag{3.13d}$$

where $\mathbf{V}_i$ is the variance of the $i^{\text{th}}$ effect.

As in the univariate case, taking $\mathbf{X} = 0$ will reduce this method to the multiple-trait counterpart of PMLE. Here, unshrunk QTL effects $\mathbf{B} = \mathbf{0}$.

## 3.6 Hypothesis test for multivariate case

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

   In the first stage, we select all QTLs with $|\mathbf{E}_j'\mathbf{V}^{-1}\mathbf{1}|/\sqrt{|\mathbf{V}|} > 10^{-6}$, where $|\mathbf{V}|$ is the determinant of matrix $\mathbf{V}$. In the second stage, we compute the likelihood ratio statistic for each QTLs with $\text{LRT}_j = -2[L(\boldsymbol{\theta}_{-j}) - L(\boldsymbol{\theta})]$, where $\boldsymbol{\theta}_{-j}$ is all parameters with $\mathbf{E}_j$ set to 0. The LOD score statistic may be used as an alternative.

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 previous derived formulas, setting the appropriate effects to zeros. The combined LOD score is defined as the minimum of the LOD scores in the tested interval.

3. **QTL by environment (QxE) 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 in different environments.

   Let $\mathbf{a}_i$ and $\mathbf{d}_i$ be the additive and dominance effects of a tested QTL at trait $i$. The hypothesis to be tested is $H_0 : \mathbf{a}_i = \mathbf{a}_j = \mathbf{a} \wedge \mathbf{d}_i = \mathbf{d}_j = \mathbf{d},\ \forall i, j \in \{1, \ldots, t\}$ *vs.* $H_A$ otherwise.

$$\begin{aligned}
\mathbf{a}_j^{(t+1)} &= \mathbf{1}_t'(\boldsymbol{\Sigma}^{(t)})^{-1}\left[\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_1 - (\mathbf{1}_t' \otimes \mathbf{1}_n)'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_2)d^{(t)} + \boldsymbol{\mu}_{\alpha_j}^{(t)}\mathbf{V}_{\alpha_j}^{-1(t)}\mathbf{V}^{(t)}\right] \\
&\quad \times \left[\left[c^{(t)}\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_1)\right]\mathbf{I}_t + \mathbf{V}_{\alpha_j}^{-1(t)}\mathbf{V}^{(t)}\right]^{-1} \\
\mathbf{d}_j^{(t+1)} &= \mathbf{1}_t'(\boldsymbol{\Sigma}^{(t)})^{-1}\left[\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(t)}\right)'\boldsymbol{\Pi}^{(t)}\mathbf{d}_2 - (\mathbf{1}_t' \otimes \mathbf{1}_n)'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_1 \# \mathbf{d}_2)a^{(t)} + \boldsymbol{\mu}_{\delta_j}^{(t)}\mathbf{V}_{\delta_j}^{-1(t)}\mathbf{V}^{(t)}\right] \\
&\quad \times \left[\left[c^{(t)}\mathbf{1}'\boldsymbol{\Pi}^{(t)}(\mathbf{d}_2 \# \mathbf{d}_2)\right]\mathbf{I}_t + \mathbf{V}_{\delta_j}^{-1(t)}\mathbf{V}^{(t)}\right]^{-1}
\end{aligned}$$

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

Under $H_0$, the log-likelihood is obtained by substitution of $\mathbf{a}$ and $\mathbf{d}$ in the $L_1$ formula with the above. The LOD score is obtained similarly.

## 3.7 Results

### 3.7.1 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.

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) as presented in Figure 3.1 (a) and (b). CIM gives narrower peaks than SIM. These figures illustrate that CIM is more precise than SIM and MIM is more precise than CIM.

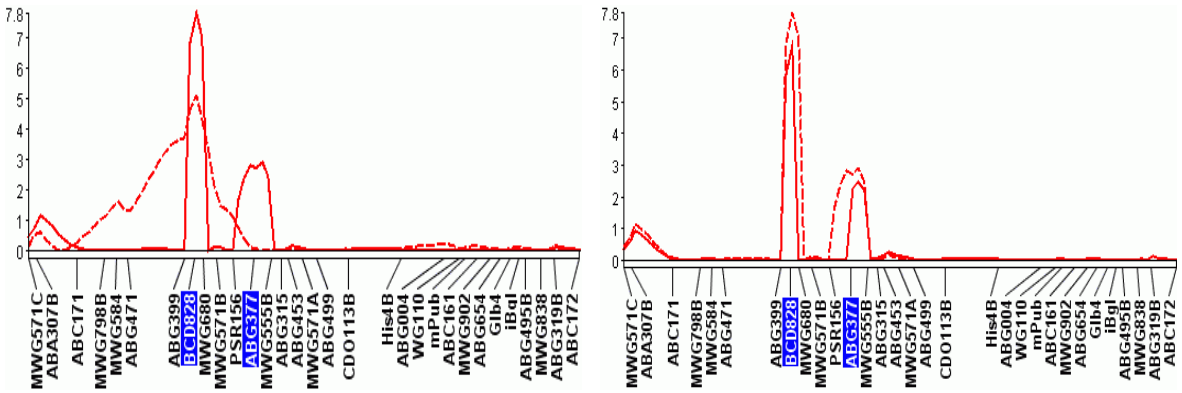
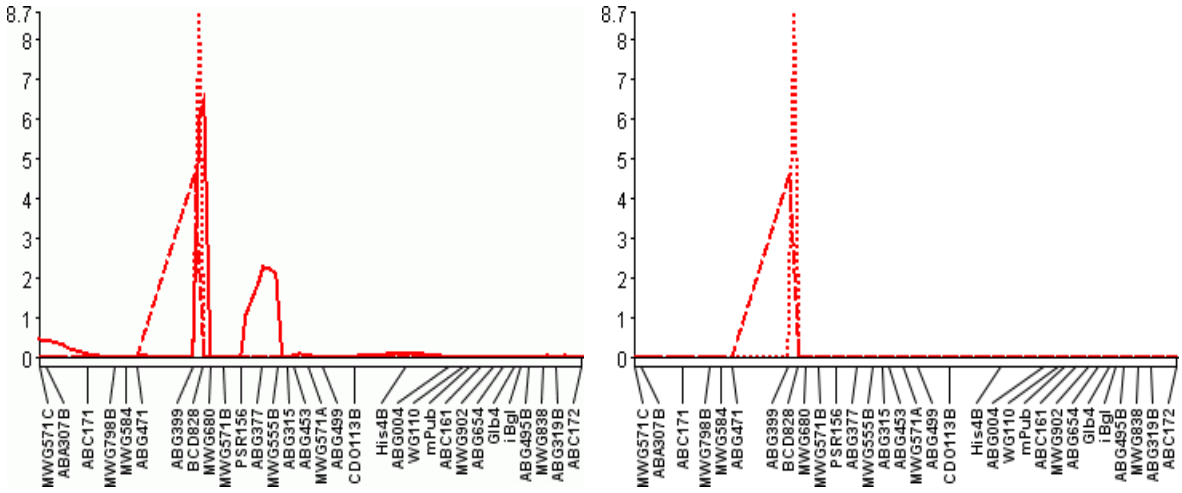



**Figure 3.1**: *Comparative QTL-detection precision of SIM, CIM, and MIM. (a) CIM improves upon SIM by giving narrower peaks. Broken line: SIM, solid line: CIM; (b) MIM improves upon CIM by giving even narrower peaks. Broken line: CIM, solid line: MIM. Highlighted markers represent cofactors selected for CIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

### 3.7.2 Results of single-trait shrinkIM on barley data

The preliminary results of shrinkIM applied to the first trait (*yld01*) of the Steptoe × Morex barley data are promising. ShrinkIM gives even narrower peaks than MIM and PMLE and yields even higher LOD scores at these peaks.




(a) (b)

**Figure 3.2**: *Comparative QTL-detection precision of MIM, PMLE, and shrinkIM. (a) For trait yld01, MIM coincides with shrinkIM. (b) Comparison between PMLE and shrinkIM. Solid line: MIM, broken line: PMLE, dotted line: shrinkIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

### 3.7.3 Simulation setup and results for single-trait shrinkIM

Two hundred progeny per dataset were generated with QGene. $F_2$ and RIL mating designs were simulated. The *heritability*, or the proportion of the phenotypic variance accounted for by genotypic variance, was set to 0.5. Two QTLs were simulated in the model, $Q_1$, fixed at 35 centiMorgans (cM) and $Q_2$ at either 55 cM or 85 cM, with no dominance effects. The additive effect of $Q_1$ was fixed at 1. The additive effect of $Q_2$ was varied between $-1$, 0.1, and 0.5 to simulate repulsion-phase QTLs (QTLs whose increasing alleles are inherited from opposite parents) and small and large coupling-phase QTLs (those whose increasing alleles are inherited from the same parent). The QTLs were placed on a 120-cM chromosome, with markers at 10-cM intervals. Scan interval was set to 2 cM. In total, 12 scenarios were

considered: 2 mapping designs, 2 QTL positions, and 3 effect values. 500 datasets were created for each scenario.

The datasets were evaluated by the MIM, PMLE, and shrinkIM methods. For MIM, a QTL was declared if a LOD score greater than 2.5 was detected within 10 cM of the QTL location. For PMLE and shrinkIM, a QTL was declared if a LOD score greater than 0.1 was detected within 10 cM of the QTL location. For MIM, the QTL width was measured from the LOD 1 drop from the QTL peak. For PMLE and shrinkIM, the QTL width was measured from the LOD 0 line. In all methods, the QTL distance was measured from the peak of the detected QTL. Both QTL width and distance were recorded only when the method could detect the QTL. The width and distance data from $Q_1$ and $Q_2$ were combined without distinction between the source QTLs.

**Table 3.1**: *Comparative QTL-detection accuracy of MIM, PMLE, and shrinkIM, from simulation. In each cell, the numbers are from MIM, PMLE, and shrinkIM, in that order. Average QTL distance is the average distance between the detected QTL peak and the actual location. Pos. stands for QTL position.*

| **Mating design** | **$Q_2$** | | **% $Q_2$ detected** | **Avg. QTL width** (cM) | **Avg. QTL distance** (cM) |
|---|---|---|---|---|---|
| | **Pos.** | **Effect** | | | |
| $F_2$ | 55 cM | 0.5 | 11.4/60.2/58.8 | 11.2/22.6/4.0 | 4.4/7.8/4.5 |
| | | 0.1 | 0/1.8/3.6 | 9.2/20.5/4.0 | 4.3/5.2/1.2 |
| | | -1 | 97.2/97.6/95.6 | 12.1/34.2/4.0 | 4.6/9.4/5.3 |
| | 85 cM | 0.5 | 92.0/89.4/93.2 | 8.1/20.0/4.0 | 4.2/5.0/3.4 |
| | | 0.1 | 0/0.8/2.4 | 8.4/20.0/4.0 | 3.8/5.2/3.3 |
| | | -1 | 98.2/99.2/98.8 | 8.8/20.0/4.0 | 4.4/5.0/3.5 |
| RIL | 55 cM | 0.5 | 18.2/58.8/61.4 | 8.1/20.0/4.0 | 3.2/5.1/3.1 |
| | | 0.1 | 0/2.2/3.0 | 8.2/20.5/4.0 | 3.9/5.1/1.4 |
| | | -1 | 95.8/97.2/97.4 | 8.1/21.0/4.0 | 5.4/7.4/2.2 |
| | 85 cM | 0.5 | 95.2/95.4/96.2 | 6.6/21.2/4.0 | 3.8/5.3/3.3 |
| | | 0.1 | 0/3.2/5.0 | 6.8/20.2/4.0 | 5.0/8.3/4.1 |
| | | -1 | 98.2/99.0/99.2 | 7.2/21.3/4.0 | 4.5/7.1/3.2 |

The simulation result is summarized in Table 3.1. All methods could always detect $Q_1$, although the QTL location might be slightly outside the peak. The peak width of shrinkIM

is always twice the scan interval since the QTL search algorithm always follows the scan interval. Likewise, the peak width of PMLE is always twice the distance between consecutive markers. When the genotypes of the marker adjacent to the flanking markers of the QTL were also almost the same (*collinear*) by chance, the peak width could be widened to three times the marker distance. This occurs only occasionally.

All methods could detect repulsion-phase QTLs. Surprisingly, these methods could not well detect QTLs with effects of the same sign, though PMLE and shrinkIM had some success. MIM could not detect QTLs with very small effects and PMLE could do so only rarely. ShrinkIM gave only modest improvement in QTL detection rate.

PMLE estimated the largest QTL effects accurately. For example, the average estimate of the additive effect of QTL 1 across all scenarios was 0.997. In contrast, shrinkIM overestimated the effect, with the average estimate being 1.174. However, PMLE severely underestimated the effect of QTLs that were of the same magnitude. For example, in scenarios where the true value of the additive effect of QTL 2 was 0.5, the average estimate from PMLE was 0.317, while the estimate from shrinkIM was 0.389. When the true value of the effect of QTL 2 was $-1.0$, the average estimate from PMLE was $-1.089$, while that of shrinkIM was $-1.117$.

PMLE is not immune to false positives, as shown in Figure 3.3, contrary to the claims made by its authors (Zhang and Xu 2005). After some inspections, the false positives were usually caused by collinearity of the marker nearby. Although shrinkIM is susceptible to false positives, its QTL location estimates are generally closer than those of PMLE.

### 3.7.4 Results of multiple-trait PMLE and shrinkIM on barley data

Multiple-trait PMLE (MT-PMLE) and shrinkIM (MT-shrinkIM) were applied to the first three traits (*yld01–yld03*) of the Steptoe × Morex barley data. The LOD plot for chromosomes 2 and 3 is shown in Figure 3.4. These multiple-trait plots follow closely the corresponding single-trait plots shown in Figure 3.5. The joint mapping plot resembles a union

of the single-trait plots. Since the estimates in single-trait shrinkIM and PMLE differ from those in single-trait MIM (*cf.* Figure 3.6), the estimates of the multiple-trait counterparts will also differ.

The QTL in chromosome 3 is known to act differently in different locations (Hayes *et al.*

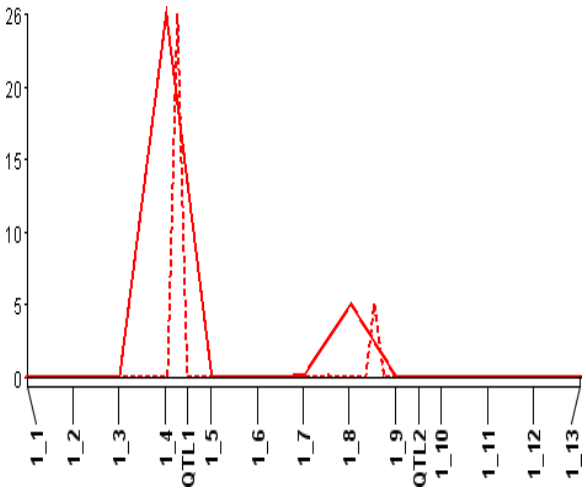



**Figure 3.3**: *PMLE and shrinkIM are not immune to false positives. The peaks to the right do not cover QTL $Q_2$. The peak given by shrinkIM is closer to the true location. Solid line: PMLE, broken line: shrinkIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

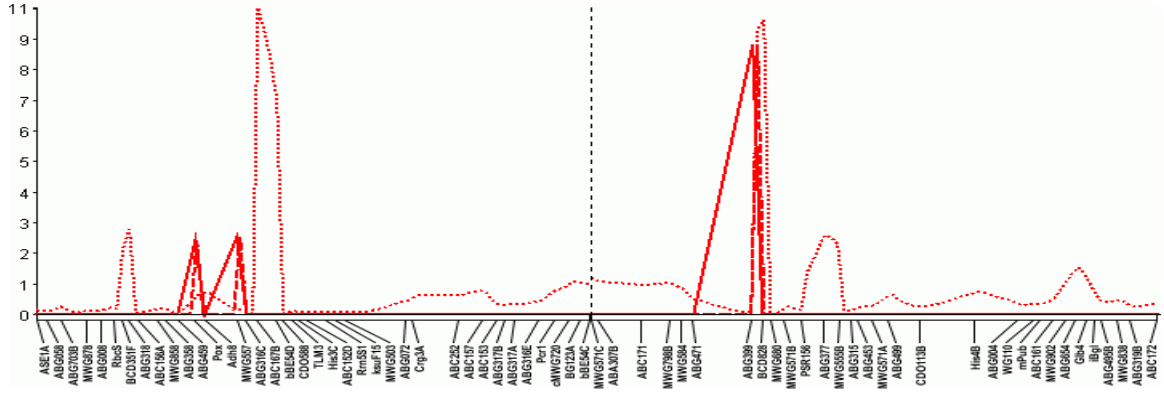



**Figure 3.4**: *Multiple-trait PMLE, shrinkIM, and MIM. The LOD profiles of multiple-trait PMLE (MT-PMLE), shrinkIM (MT-shrinkIM), and MIM (MT-MIM) for traits* yld01–yld03 *on chromosomes 2 and 3 of Steptoe × Morex dataset. Solid line: MT-PMLE, broken line: MT-shrinkIM, dotted line: MT-MIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

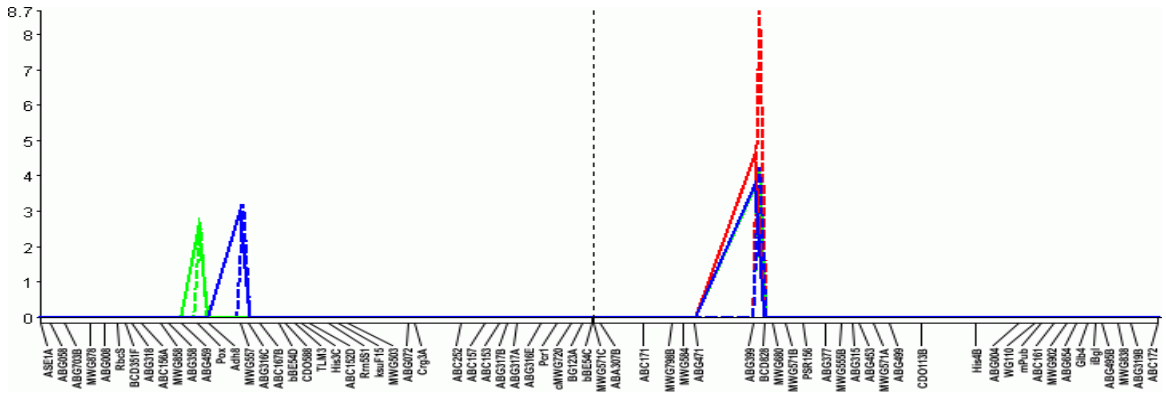



**Figure 3.5**: *The LOD profiles of single-trait PMLE and shrinkIM for traits* yld01–yld03 *on chromosomes 2 and 3 of Steptoe × Morex dataset. Solid line: MT-PMLE, broken line: MT-shrinkIM, dotted line: MT-MIM. Colors correspond to traits. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

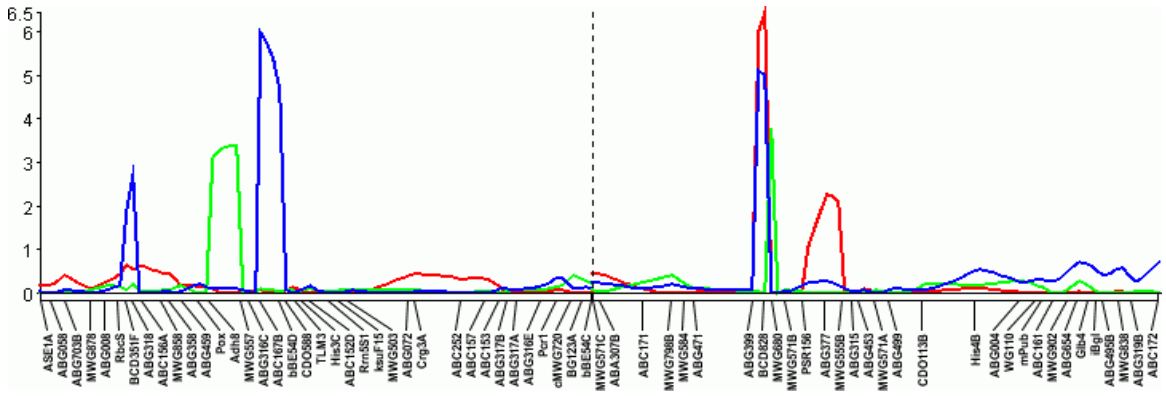



**Figure 3.6**: *The LOD profiles of single-trait MIM for traits* yld01–yld03 *on chromosomes 2 and 3 of Steptoe × Morex dataset. Colors correspond to traits. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

1994). MT-PMLE and MT-shrinkIM could detect them, as shown in Figure 3.7. However the peak for the Q×E test did not pass the permutation test.

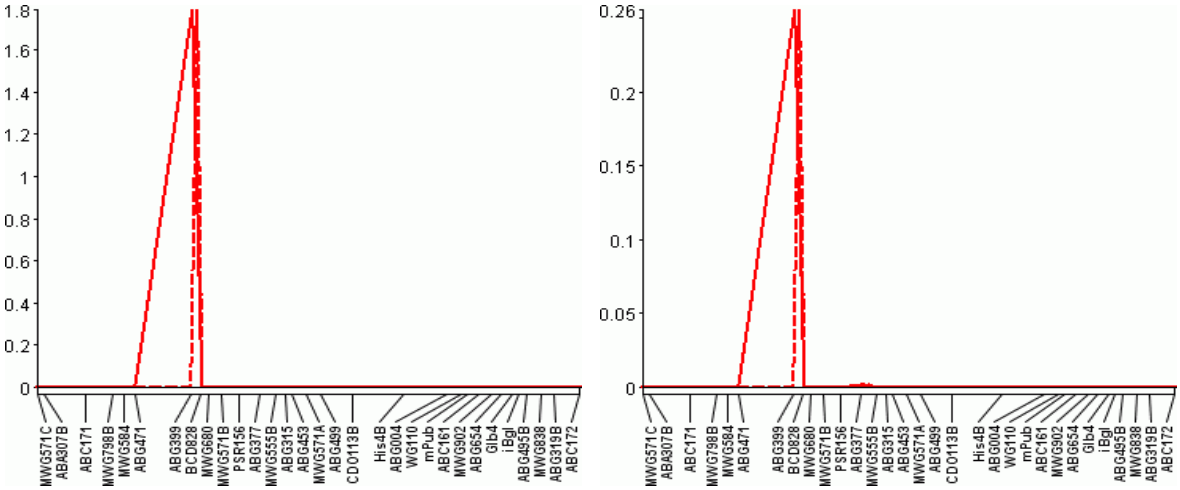



**Figure 3.7**: *Pleiotropy and Q×E tests with PMLE and shrinkIM. (a) Pleiotropy analysis of traits* yld01–yld03 *of Steptoe × Morex dataset; (b) Q×E analysis of the same traits. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome.*

### 3.7.5 Simulation setup and results for multiple-trait shrinkIM

Five hundred simulated datasets of 200 progeny each, as in the single-trait study, were generated with 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. For each trait there were two QTLs in the model, $Q_1$ at 35 centiMorgans (cM) and $Q_2$ at 55 or 85 cM, with the effects described in Table 3.2. The QTLs were placed on a 120-cM chromosome, with markers at 10-cM intervals. Scan interval was set to 2 cM.

The datasets were evaluated by multiple-trait MIM, PMLE, and shrinkIM. For MT-MIM, a QTL was declared if a LOD score greater than 2.5 was detected within 10 cM of the QTL location. For MT-PMLE and MT-shrinkIM, a QTL was declared if a LOD score greater than 0.1 was detected within 10 cM of the QTL location. For MT-MIM, the QTL width was measured at the LOD 1 drop from the QTL peak. For MT-PMLE and

**Table 3.2**: *QTL effect values from comparative simulation study of MT-PMLE and MT-shrinkIM*

| Trait | $Q_1$ effect | | $Q_2$ effect | |
|---|---|---|---|---|
| | **Additive** | **Dominance** | **Additive** | **Dominance** |
| Trait 1 | 1.0 | 0.0 | 0.3 | 0.0 |
| Trait 2 | 0.3 | 0.0 | 0.7 | 0.0 |

MT-shrinkIM, the QTL width was measured at the LOD 0 line. In all methods, the QTL distance was measured from the peak of the detected QTL. Both QTL width and distance were recorded only when the method could detect the QTL. The width and distance data from $Q_1$ and $Q_2$ were combined without distinction between the source QTLs.

All methods could detect all QTLs when QTL 2 was placed at 85 cM. When QTL 2 was placed at 55 cM, MT-PMLE and MT-shrinkIM could detect all QTLs, whereas MT-MIM could detect only QTL $Q_1$. The QTL $Q_2$ detection rates for MT-MIM when QTL 2 was placed at 55 cM were 0.832, 0.816, and 0.818 for $\rho = 0.0, 0.4$, and 0.9. Although MT-MIM could detect QTL 2 at the lenient LOD threshold of 2.5, in many cases spurious peaks were also present. QTL width and distance were close to those of single-trait simulations.

## 3.8 Discussion and conclusion

PMLE and shrinkIM showed some potential as emerging QTL analysis methods. Both were much more precise than MIM in detecting tightly linked QTLs. ShrinkIM yields much narrower peaks that are closer to the true QTL position than PMLE because it searches over QTL scan intervals while PMLE searches over marker locations.

ShrinkIM showed marginal improvement over PMLE in detecting QTLs with weak effects. However, both methods are unable in general to detect weak QTLs. ShrinkIM tends to overestimate the largest QTL effect, although it estimates better than PMLE QTLs of lesser magnitude.

The multiple-trait extension of PMLE and shrinkIM produced clean LOD profiles, as

did their single-trait counterparts. The simulation results suggested that MT-PMLE and MT-shrinkIM could exploit correlation structure to improve detection of linked QTLs. MT-MIM, using the same strategy, showed comparable success but with more spurious peaks for QTLs of small effect.

# Chapter 4

# QTL mapping methods based on the generalized linear model


## Abstract

Statistical methods currently available for mapping QTLs are generally sensitive to some degree of non-normality in the trait data, whereas some economically important traits such as plant disease resistance are known to be non-normally distributed. In addition, since some traits are non-numeric, such standard methods may not be applicable. Methods have been developed to overcome instances of this problem, some of which, such as (Wittenburg *et al.* 2007; Hackett and Weller 1995), are based on generalized linear model (GLZ). I describe a more general framework based on GLZ, developed to map QTLs for non-normal and non-numeric trait data. I also describe the multivariate extension to handle polytomous (multicategorical) trait data. I tested this method with simulation in four scenarios: traits with skewed distribution, ordinal and multicategorical traits, and traits governed by XOR logic.


## 4.1 Introduction

Some economically important traits, such as plant disease resistance controlled by two or more genes, are non-normally distributed. Breeders prefer partial to complete resistance genes given a disease since complete resistance genes can be easily overcome by mutations

of pathogen over several generations. For this reason, screening for partial resistance may be impossible because genes for such resistance is obscured by those of complete resistance. In this case, the histogram of resistance score will be skewed to the right, with most progeny showing little to no sign of infection.

Many statistical methods currently available for mapping QTLs controlling polygenic traits, such as single (SIM) (Haley and Knott 1992; Lander and Botstein 1989), and composite (CIM) (Jansen 1994; Jansen and Stam 1994; Zeng 1994), multiple (MIM) (Kao *et al.* 1999), and Bayesian interval mapping (BIM) (Satagopan *et al.* 1996; Sillanpää and Arjas 1998; Stephens and Fisch 1998) assume that trait data $\mathbf{y}$ are normally distributed. Methods that assume normality are generally sensitive to some degree of non-normality (Box 1953). Hence, when such methods are applied to non-normal data, their power is reduced. Even if there is a transformation function $f(\cdot)$ such that $f(\mathbf{y})$ is approximately normal, the QTL effect estimates may not be correct since in general $E[f(\mathbf{y})] \neq f(E[\mathbf{y}])$.

Aside from a need of mapping non-normal traits, there is a need to map QTLs controlling non-metric traits. Methods have been developed for binary (Yi and Xu 2000; Xu *et al.* 2003; Coffman *et al.* 2005; Zhu *et al.* 2007), categorical (Rao and Li 2000), and ordinal trait data (Li *et al.* 2006; Yi *et al.* 2007; Hackett and Weller 1995). Although it is possible to analyze each data type separately, it is impractical to invent new methods for each.

I seek to establish a generalization of a QTL-mapping method based on the generalized linear model (GLZ) framework (Nelder and Wedderburn 1972) for non-normal and non-metric traits. Models based on the GLZ can accommodate normally distributed traits as a special case. A method with similar approach has been developed (Wittenburg *et al.* 2007), but it was limited to gamma-distributed trait data for sample variance estimation. The model I develop here accommodates any distribution in the exponential family.

## 4.2 Method

The general linear model (GLM) for genetic mapping of a quantitative trait is formulated as

$$y_j = \mu_0 + x_j a + z_j d + \sum_{k=1}^{p} c_{jk}\beta_k^* + \epsilon_j,$$

where $y_j$ is the trait value for offspring $j$; $\mu_0$ is the general mean; $a$ and $d$ are additive and dominant effects for the QTL; $\beta_k^*$ is the fixed effect of covariate $k$; $x_j$ is 1 if the genotype of the QTL of offspring $j$ is $AA$, $-1$ if $aa$, and 0 otherwise; $z_j$ is $\frac{1}{2}$ if the QTL genotype is heterozygous, $-\frac{1}{2}$ otherwise; $c_{jk}$ is the value of covariate $k$ of offspring $j$; and $\epsilon_j$ is the error term, where $\epsilon_j \sim N(0,\ \sigma^2)$. This model can be written in matrix form as $\mathbf{y} = \mathbf{1}\mu_0 + \mathbf{x}a + \mathbf{z}d + \mathbf{C}\boldsymbol{\beta}^* + \mathbf{e}$, or more compactly, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y}$ is an $n \times 1$ column vector, $\mathbf{X}$ is an $n \times (3+p)$ design matrix, $\boldsymbol{\beta}$ is an $(3+p) \times 1$ column vector of effects, and $\mathbf{e} \sim N(0, \sigma^2\mathbf{I}_n)$, with $\mathbf{I}_n$ is an $n \times n$ identity matrix. It follows that $E[\mathbf{y}] = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

In GLZ, the trait distribution may be any member of an exponential family. A link function $g(\cdot)$, an invertible function that transforms the systematic component $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, is added to the model so that $E[\mathbf{y}] = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = g^{-1}(\boldsymbol{\eta})$. The choice of link function depends on the trait distribution and the relationship between the variance and the mean. Common distributions used in GLZ, typical link functions, variances, and uses are summarized in Table 4.1.

The pdf or pmf used in the model is derived from the exponential dispersion family (Jørgensen 1983) and is parameterized as follows:

$$f(y_j; \theta_j, \phi) = \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j, \phi)\right], \tag{4.1}$$

where $\phi$ is the dispersion parameter and $\theta_j$ is the *natural parameter* of the distribution. Functions $a(\phi)$, $b(\theta_j)$, and $c(y_j, \phi)$ are replaced with suitable forms for each distribution in the exponential family. For example, for a Poisson distribution with mean $\lambda_j$, $a(\phi) = 1$, $b(\theta_j) = \exp(\theta_j)$, and $c(y_j, \phi) = -\log y_j!$, while $\theta_j = \log \lambda_j$. When $\phi$ is known, Equation 4.1

reduces to the exponential family pdf or pmf, which is

$$f(y_j;\theta_j) = h(y_j)\exp\left[e(\theta_j)T(y_j) - A(\theta_j)\right]. \tag{4.2}$$

Write $h(y_j) = \exp[c(y_j,\phi)]$, $T(y_j) = y_j$, $e(\theta_j) = \theta_j/a(\phi)$, and $A(\theta_j) = b(\theta_j)/a(\phi)$.

The solution can be obtained by maximum-likelihood estimation (Agresti 2002, p. 133). Let $l_j$ be the log-likelihood function of offspring $j$, *i.e.*,

$$l_j = \frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j,\phi)$$

Differentiation with respect to $\theta_j$ yields $\mu_j = \mathrm{E}[Y_j] = b'(\theta_j)$, where $b'(\theta_j) = \partial b(\theta_j)/\partial\theta_j$ and $\mathrm{Var}[Y_j] = b''(\theta_j)a(\phi)$, where $b''(\theta_j) = \partial^2 b(\theta_j)/\partial\theta_j^2$.

The joint log likelihood of $\boldsymbol{\beta} = (a, d, \boldsymbol{\beta}^*)$ for $n$ offspring is

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{n} \frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j,\phi)$$

where $\boldsymbol{\beta}$ depends on $\boldsymbol{\theta}$.

The likelihood equations are

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_i} = \sum_{j=1}^{n} \frac{\partial l_j}{\partial\beta_i}$$

**Table 4.1**: *Typical error term distributions in the generalized linear model, with associated link function and variance. Variance is expressed in terms of the mean,* $\mu$. $\phi$ *is a dispersion parameter.*

| **Distribution** | **Common link functions** | **Variance** | **Typical use** |
|---|---|---|---|
| Binomial | Logit, probit | $\mu(1-\mu)$ | Binary data |
| Binomial | Complementary log-log | $\mu(1-\mu)$ | Survival analysis |
| Normal | Identity | $\mu$ | Normally distributed data |
| Poisson | Log | $\mu$ | Count data |
| Negative-binomial | Log | $\frac{\mu^2}{\phi} + \mu$ | Count data with dispersion |
| Gamma | Reciprocal | $\mu^2$ | Estimating sample variance |
| Inverse Gaussian | Reciprocal squared | $\mu^3$ | Positively skewed data |

The chain rule is used to differentiate the log likelihood, as follows:

$$\frac{\partial l_j}{\partial \beta_i} = \frac{\partial l_j}{\partial \theta_j}\frac{\partial \theta_j}{\partial \mu_j}\frac{\partial \mu_j}{\partial \eta_j}\frac{\partial \eta_j}{\partial \beta_i}$$

Since $\eta_j = x_j a + z_j d + \sum_{k=1}^{p} c_{jk}\beta_k^*$ and $\mu_j = g^{-1}(\eta_j)$, the derivation for the terms are

$$\frac{\partial l_j}{\partial \theta_j} = \frac{y_j - \mu_j}{a(\phi)} \qquad \frac{\mu_j}{\theta_j} = b''(\theta_j) = \frac{\mathrm{Var}[Y_j]}{a(\phi)}$$

$$\frac{\partial \eta_j}{\partial a} = x_j \qquad \frac{\partial \eta_j}{\partial d} = z_j \qquad \frac{\partial \eta_j}{\partial \beta_k^*} = c_{jk}$$

$$\frac{\partial \mu_j}{\partial \eta_j} = \frac{1}{g'(g^{-1}(\eta_j))} = \frac{1}{g'(\mu_j)}$$

Thus, the likelihood equations are

$$\sum_{j=1}^{n} \frac{\partial l_j}{\partial \beta_i} = \frac{y_j - \mu_j}{a(\phi)}\frac{a(\phi)}{\mathrm{Var}[Y_j]}\frac{1}{g'(\mu_j)}x_{ij}^* = \frac{(y_j - \mu_j)x_{ij}^*}{\mathrm{Var}[Y_j]g'(\mu_j)} = 0$$

where

$$x_{ij}^* = \begin{cases} x_j & \text{if } \beta_i = a \\ z_j & \text{if } \beta_i = j \\ c_{jk} & \text{if } \beta_i = \beta_k^* \end{cases}$$

The iteratively-reweighted least-squares method is used to solve these equations (Agresti 2002, pp. 146–147). In this method, the Fisher scores are calculated to find ML estimates. Let $\boldsymbol{\eta} = (\eta_j)', j \in \{1 \ldots n\}$. The solution is

$$\begin{aligned}
\boldsymbol{\eta}_1 &= \mathbf{x}a^{(t)} + \mathbf{z}d^{(t)} + \mathbf{C}\boldsymbol{\beta}^{*(t)} \\
a^{(t+1)} &= \frac{\mathbf{x}'\mathbf{W}^{(t)}}{\mathbf{x}'\mathbf{W}^{(t)}\mathbf{x}}\left[\boldsymbol{\eta}_1 + (\mathbf{y} - \mathbf{z}d^{(t)} - \mathbf{C}\boldsymbol{\beta}^{*(t)} - g^{-1}(\boldsymbol{\eta}_1))D(\boldsymbol{\eta}_1)^{-1}\right] \\
\boldsymbol{\eta}_2 &= \mathbf{x}a^{(t+1)} + \mathbf{z}d^{(t)} + \mathbf{C}\boldsymbol{\beta}^{*(t)} \\
d^{(t+1)} &= \frac{\mathbf{z}'\mathbf{W}^{(t)}}{\mathbf{z}'\mathbf{W}^{(t)}\mathbf{z}}\left[\boldsymbol{\eta}_2 + (\mathbf{y} - \mathbf{x}a^{(t+1)} - \mathbf{C}\boldsymbol{\beta}^{*(t)} - g^{-1}(\boldsymbol{\eta}_2))D(\boldsymbol{\eta}_2)^{-1}\right] \\
\boldsymbol{\eta}_3 &= \mathbf{x}a^{(t+1)} + \mathbf{z}d^{(t+1)} + \mathbf{C}\boldsymbol{\beta}^{*(t)} \\
\boldsymbol{\beta}^{*(t+1)} &= (\mathbf{C}'\mathbf{W}^{(t)}\mathbf{C})^{-1}\mathbf{C}'\mathbf{W}^{(t)}\left[\boldsymbol{\eta}_3 + (\mathbf{y} - \mathbf{x}a^{(t+1)} - \mathbf{z}d^{(t+1)} - g^{-1}(\boldsymbol{\eta}_3))D(\boldsymbol{\eta}_3)^{-1}\right]
\end{aligned}$$

where $\mathbf{W}$ is a diagonal matrix with $w_{jj} = (\partial\mu_j/\partial\eta_j)^2/\mathrm{Var}[Y_i]$ and $D(\mathbf{x})$ is an $n \times 1$ column vector where $D(x_i) = \partial g^{-1}(x_i)/\partial x_i$. Initial estimates can be obtained by the least-squares method.

Let $\hat{\boldsymbol{\eta}}$ be the estimated systematic component obtained by the preceding formulas. The fitted value, $\hat{\boldsymbol{\mu}}$, is $g^{-1}(\hat{\boldsymbol{\eta}})$. The maximum likelihood of the model is therefore $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$. The maximum likelihood for the saturated model is $l(\mu; \mathbf{y})$. The difference of *deviance* between two successive iterations can be used as the termination condition, where deviance is defined as $\mathrm{Dev} = -2[l(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l(\mu; \mathbf{y})]$. Since the deviance is a likelihood-ratio test, it can be converted to a LOD score by $\mathrm{LOD} = \mathrm{Dev}/(2 \log 10)$.

This solution can be applied to an interval-mapping setting, with $\mathbf{x}$ and $\mathbf{z}$ defined as in simple interval mapping (SIM). For a composite IM extension, cofactor marker genotypes can be incorporated in matrix $\mathbf{C}$.

## 4.3 MIM-based extension

The model discussed in the previous section can be extended to a multiple-QTL model as follows:

$$y_j = \mu_0 + \sum_{i=1}^{q}(x_{ji}a_i + z_{ji}d_i) + \sum_{k=1}^{p} c_{jk}\beta_k^* + \epsilon_j,$$

The derivation and the solution are similar to those presented in the previous section. The only difference in the solution is that each of the multiple QTL effects is updated in sequence. For each QTL $i$, update

$$\begin{aligned}
\boldsymbol{\eta}_{i1} &= \sum_{h=1}^{i-1}\left(\mathbf{x}_h a_h^{(t+1)} + \mathbf{z}_h d_h^{(t+1)}\right) + \sum_{h=i}^{q}\left(\mathbf{x}_h a_h^{(t)} + \mathbf{z}_h d_h^{(t)}\right) + \mathbf{C}\boldsymbol{\beta}^{*(t)} \\
a_i^{(t+1)} &= \frac{\mathbf{x}_i'\mathbf{W}^{(t)}}{\mathbf{x}_i'\mathbf{W}^{(t)}\mathbf{x}_i}\left[\boldsymbol{\eta}_{i1} + (\mathbf{y} - \boldsymbol{\eta}_{i1} + \mathbf{x}_i a_i^{(t)} - g^{-1}(\boldsymbol{\eta}_{i1}))D(\boldsymbol{\eta}_{i1})^{-1}\right] \\
\boldsymbol{\eta}_{i2} &= \sum_{h=1}^{i-1}\left(\mathbf{x}_h a_h^{(t+1)} + \mathbf{z}_h d_h^{(t+1)}\right) + \mathbf{x}_h a_h^{(t+1)} + \mathbf{z}_h d_h^{(t)} + \sum_{h=i+1}^{q}\left(\mathbf{x}_h a_h^{(t)} + \mathbf{z}_h d_h^{(t)}\right) + \mathbf{C}\boldsymbol{\beta}^{*(t)} \\
d_i^{(t+1)} &= \frac{\mathbf{z}_i'\mathbf{W}^{(t)}}{\mathbf{z}_i'\mathbf{W}^{(t)}\mathbf{z}_i}\left[\boldsymbol{\eta}_{i2} + (\mathbf{y} - \boldsymbol{\eta}_{i2} + \mathbf{z}_i d_i^{(t)} - g^{-1}(\boldsymbol{\eta}_{i2}))D(\boldsymbol{\eta}_{i2})^{-1}\right]
\end{aligned}$$

followed by

$$\boldsymbol{\eta}_3 = \sum_{h=1}^{q} \left( \mathbf{x}_h a_h^{(t+1)} + \mathbf{z}_h d_h^{(t+1)} \right) + \mathbf{C}\boldsymbol{\beta}^{*(t)}$$
$$\boldsymbol{\beta}^{*(t+1)} = (\mathbf{C}'\mathbf{W}^{(t)}\mathbf{C})^{-1}\mathbf{C}'\mathbf{W}^{(t)} \left[ \boldsymbol{\eta}_3 + (\mathbf{y} - \boldsymbol{\eta}_{i3} + \mathbf{C}\boldsymbol{\beta}^{(t)} - g^{-1}(\boldsymbol{\eta}_3))D(\boldsymbol{\eta}_3)^{-1} \right]$$

Initial estimates can be obtained by the least-squares method, as in the previous section. The difference of deviance between two successive iterations can be used as the termination condition.

## 4.4 Univariate implementation

The QTL and covariate effects can be updated at the same time instead of sequentially as shown in the previous sections. Write $\boldsymbol{\beta} = (a_1|d_1|\ldots|a_q|d_q|\boldsymbol{\beta}^{*'})'$ and $\mathbf{X} = (\mathbf{x}_1|\mathbf{z}_1|\ldots|\mathbf{x}_q|\mathbf{z}_q|\mathbf{C})$. The solution can be written as:

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{u}$$

where $\mathbf{u}$ is an $n \times 1$ column vector where

$$u_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

Let $\mathbf{W}^{1/2}$ be the Cholesky decomposition of $\mathbf{W}$. Since $\mathbf{W}$ is a diagonal matrix with elements $w_{jj}$, its Cholesky decomposition is also a diagonal matrix with elements $\sqrt{w_{jj}}$.

Write $\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X}$ and $\mathbf{u}^* = \mathbf{W}^{1/2}\mathbf{u}$ to yield

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{u}^*$$

This is the least-squares solution of $\mathbf{u}^*$ given $\mathbf{X}^*$. Now a QR decomposition can be performed on $\mathbf{X}^* = \mathbf{Q}\mathbf{R}$ to yield

$$\mathbf{R}\boldsymbol{\beta}^{(t+1)} = \mathbf{Q}'\mathbf{u}^*$$

With the partial QR updating described in §2.2, the computation can be performed efficiently. QR updating can also be applied similarly to the sequential-updating scenario.

## 4.5 Multivariate extension

The GLZ model for multiple QTL can be extended for multiple traits. The model for $t$ traits and $q$ QTLs is

$$\mathbf{y}_j = \mu_0\mathbf{1} + \sum_{i=1}^{q}(x_{ji}\mathbf{a}_i + z_{ji}\mathbf{d}_i) + \sum_{k=1}^{p} c_{jk}\boldsymbol{\beta}_k^* + \boldsymbol{\epsilon}_j.$$

The variables are the same as in the original model except that $\mathbf{y}_j$, $\mathbf{a}_i$, $\mathbf{d}_i$, and $\boldsymbol{\beta}_k^*$ are $t \times 1$ column vectors instead of scalars.

Let $\mathbf{B}^* = (\boldsymbol{\beta}_1^* | \ldots | \boldsymbol{\beta}_p^*)'$ be a $p \times t$ matrix of covariate effects and $\mathbf{Y}$ be an $n \times t$ trait data matrix. For each QTL $i$, update:

$$\boldsymbol{\eta}_{i1} = \sum_{h=1}^{i-1}\left(\mathbf{x}_h\mathbf{a}_h^{'(t+1)} + \mathbf{z}_h\mathbf{d}_h^{'(t+1)}\right) + \sum_{h=i}^{q}\left(\mathbf{x}_h\mathbf{a}_h^{(t)} + \mathbf{z}_h\mathbf{d}_h^{(t)}\right) + \mathbf{C}\mathbf{B}^{*(t)}$$

$$\underset{n\times t}{\mathbf{S}}, \text{ such that } \underset{1\times t}{\mathbf{s_j}} = \left[\boldsymbol{\eta}_{i1_j} + (\mathbf{y_j} - \boldsymbol{\eta}_{i1_j} + \mathbf{x}_i\mathbf{a}_i^{'(t)} - g^{-1}(\boldsymbol{\eta}_{i1_j}))D(\boldsymbol{\eta}_{i1_j})^{-1}\right]$$

$$\mathbf{a}_i^{'(t+1)} = \left[(\mathbf{J}_{t,t}\otimes\mathbf{x}_i')\mathbf{W}^{(t)}(\mathbf{x}_i\otimes\mathbf{J}_{t,t})\right]^{-1}(\mathbf{J}_{t,t}\otimes\mathbf{x}_i')\mathbf{W}^{(t)}\mathbf{vec}(\mathbf{S}')$$

$$\boldsymbol{\eta}_{i2} = \sum_{h=1}^{i-1}\left(\mathbf{x}_h\mathbf{a}_h^{'(t+1)} + \mathbf{z}_h\mathbf{d}_h^{'(t+1)}\right) + \mathbf{x}_h\mathbf{a}_h^{'(t+1)} + \mathbf{z}_h\mathbf{d}_h^{'(t)} + \sum_{h=i+1}^{q}\left(\mathbf{x}_h\mathbf{a}_h^{'(t)} + \mathbf{z}_h\mathbf{d}_h^{'(t)}\right) + \mathbf{C}\mathbf{B}^{*(t)}$$

$$\underset{n\times t}{\mathbf{T}}, \text{ such that } \underset{1\times t}{\mathbf{t_j}} = \left[\boldsymbol{\eta}_{i2_j} + (\mathbf{y_j} - \boldsymbol{\eta}_{i2_j} + \mathbf{z}_i\mathbf{d}_i^{'(t)} - g^{-1}(\boldsymbol{\eta}_{i2_j}))D(\boldsymbol{\eta}_{i2_j})^{-1}\right]$$

$$\mathbf{d}_i^{'(t+1)} = \left[(\mathbf{J}_{t,t}\otimes\mathbf{z}_i')\mathbf{W}^{(t)}(\mathbf{z}_i\otimes\mathbf{J}_{t,t})\right]^{-1}(\mathbf{J}_{t,t}\otimes\mathbf{z}_i')\mathbf{W}^{(t)}\mathbf{vec}(\mathbf{T}')$$

followed by

$$\boldsymbol{\eta}_3 = \sum_{h=1}^{q}\left(\mathbf{x}_h\mathbf{a}_h^{'(t+1)} + \mathbf{z}_h\mathbf{d}_h^{'(t+1)}\right) + \mathbf{C}\mathbf{B}^{*(t)}$$

$$\underset{n\times t}{\mathbf{U}}, \text{ such that } \underset{1\times t}{\mathbf{u_j}} = \left[\boldsymbol{\eta}_{i3_j} + (\mathbf{y_j} - \boldsymbol{\eta}_{i3_j} + \mathbf{C}\mathbf{B}^{*(t)} - g^{-1}(\boldsymbol{\eta}_{i3_j}))D(\boldsymbol{\eta}_{i3_j})^{-1}\right]$$

$$\mathbf{vec}(\mathbf{B}^{*'(t+1)}) = ((\mathbf{J}_{t,t}\otimes\mathbf{C}')\mathbf{W}^{(t)}(\mathbf{C}\otimes\mathbf{C}))^{-1}(\mathbf{J}_{t,t}\otimes\mathbf{C}')\mathbf{W}^{(t)}\mathbf{vec}(\mathbf{U}')$$

where **vec** denotes the matrix vectorization, $\otimes$ the Kronecker product, and $\mathbf{J}_{p,q}$ a $p \times q$ matrix of ones.

Here $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, and $\boldsymbol{\eta}_3$ are $n \times t$ matrices. Function $D(\mathbf{x})$ takes a $t \times 1$ column vector

argument and returns a $t \times t$ matrix where

$$D(\mathbf{x}) = \frac{\partial g^{-1}(\mathbf{x})}{\partial \mathbf{x}}$$

Matrix $\mathbf{W}$ is an $nt \times nt$ block-diagonal matrix of $t \times t$, which means that if $\mathbf{W}$ is cut into $n \times n$ blocks of size $t \times t$ each, non-diagonal blocks are zeroes. Let $\mathbf{W}_{ij}$ denote the $(i,j)^{\text{th}}$ block and $\mathbf{X}_j$ the $j^{\text{th}}$ row of matrix $\mathbf{X} = (\mathbf{x_1}|\mathbf{z_1}|\ldots|\mathbf{x_q}|\mathbf{z_q}|\mathbf{C})$. Then

$$\mathbf{W}_{jj} = D(\mathbf{X}_j)'\text{Var}[\mathbf{Y}_j]D(\mathbf{X}_j)$$

**Table 4.2**: *Multivariate error term distributions and link functions used in generalized linear models.*

| **Distribution** | **Common link functions** | **Univariate dist. analog** | **Typical use** |
|---|---|---|---|
| Multinomial | Logit | Binomial | Multiple-categoricy data |
| Multinomial | Cumulative logit | Binomial | Ordinal data |
| Multinomial | Cumulative probit | Binomial | Ordinal data |
| Multivariate Normal | Identity | Univariate normal | Multivariate Normal data |
| Wishart | Inverse | Gamma | Estimating sample variance |

Multiple-category (polytomous) and ordinal data can be modeled as a multivariate GLZ with multinomial distribution (Fahmeir and Tutz 2001). For polytomous data, the multinomial logit link function is used. For ordinal data, a cumulative link function, one that transforms the systematic component $\boldsymbol{\eta}$ into cumulative probabilities, is used. For this application, the most popular choice is the cumulative logit link function.

Polytomous data with $t+1$ classes can be encoded as $t$ binary traits. The class with the most observations is omitted in the encoding.

## 4.6 Multivariate implementation

The QR decomposition optimization in the univariate implementation can be applied similarly to the multivariate one. The major difference is that the Cholesky decomposition,

instead of a simple square root, must be calculated on the diagonal block of $\mathbf{W}_{jj}$. Let $\mathbf{W}^{-1/2}$ be the Cholesky decomposition of $\mathbf{W}$. The result is also a block diagonal matrix with block elements $\mathbf{W}_{jj}^{-1/2} = \text{Chol}[\text{Var}(\mathbf{Y}_j)]D(\mathbf{X}_j)'$, where Chol[$\mathbf{X}$] is the Cholesky decomposition of matrix $\mathbf{X}$.

Another optimization can be performed on a redundant structure produced by the Kronecker products. The matrix multiplication of the sparse structure of $\mathbf{W}$ matrix with the $\mathbf{X}$ can be greatly simplified by execution one $t \times t$ block at a time, skipping the non-diagonal blocks of $\mathbf{W}$.

## 4.7 Simulation setup

Five hundred simulated datasets of 200 progeny each were generated with QGene for each case 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. One trait was generated per datafile. For each scenario there were two QTLs in the model, $Q_1$ at 35 centiMorgans (cM) and $Q_2$ at 85 cM, with additive effects of 1 and $-1$ and no dominance effects. The QTLs were placed on a 120-cM chromosome, with markers at 10-cM intervals. Scan interval was set to 2 cM.

Cases investigated in the simulation included:

- Skewed error distribution. For this purpose, a Gamma distribution with shape $\alpha_i$ and rate $\beta_i$ was used to simulate the trait value of offspring $i$. $\alpha_i$ and $\beta_i$ were chosen such that $\frac{\alpha_i}{\beta_i} = \mu_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$ and $\frac{\alpha_i}{\beta_i^2} = (1-h)s^2$, where $h$ is the heritability and $s^2$ is the sample variance of the QTL effects. The link function used here was the reciprocal function, *i.e.*, $g(x) = \frac{1}{x}$. The additive effect of $Q_2$ was set to 0.5 instead. A Gamma distribution was also used in the estimation.

- Discretization. Continuous trait values were converted into $k = 4$ integer values. There were $k-1$ threshold values, $[t_1 \leq t_2 \leq \ldots t_{k-1}]$. In this simulation, $t_i =< -2, 0, 2 >$. Values less than $t_1$ were converted to 1, with those between $t_i$ and $t_{i+1}$ converted to $i+1$. The multinomial distribution was used with a cumulative link function.

- Multicategorization. Continuous trait values were converted into $k = 4$ categories. The conversion logic was almost the same as for discretization, except the following function was used instead:

$$f(y_i) = \begin{cases} 2 & \text{if } y_i < -2 \\ 4 & \text{if } -2 \leq y_i < 0 \\ 1 & \text{if } 0 \leq y_i < 2 \\ 3 & \text{if } y_i \geq 2 \end{cases}$$

  Multinomial distribution and multinomial logit link function were used.

- XOR logic. Consider two QTLs, $A$ and $B$, with two possible genotypes each, $AA$, $aa$ and $BB$, $bb$. Here the trait is binary. Progeny that have the $AAbb$ or $aaBB$ combination will have the trait value of 1, all others 0. In computer science, this is the well-known *nonlinear separability* problem, *i.e.*, a classification problem where labels (in this case, denoting the levels of the binary trait) that lie in a $n$-dimensional hyperspace of independent variables (in this case, the QTL genotypes) cannot be separated by any single $k < n$-dimensional linear hyperplane. The logit link function and binomial distribution were used for this case. A RIL mating design was used instead of an $F_2$ and the QTL effects were ignored.

A QTL was declared if there was a LOD score of at least 2.8 within 15 cM of the true QTL position. The methods described here were compared to composite (CIM) and multiple interval mapping (MIM). The CIM counterpart of the GLZ method is labelled GLZ-CIM, while the MIM counterpart is labeled GLZ-MIM.

## 4.8 Results

### 4.8.1 Skewed trait distribution case

A typical plot of a skewed trait distribution generated by simulation is shown in Figure 4.1. Most trait values are between 0 and 0.5, while a few may be as high as 5.7. Even after a log transform, the trait distribution is still far from normally distributed (Figure 4.2), as shown

by a QQ plot, in which the normalized trait values are plotted against quantile values in the standard normal distribution.

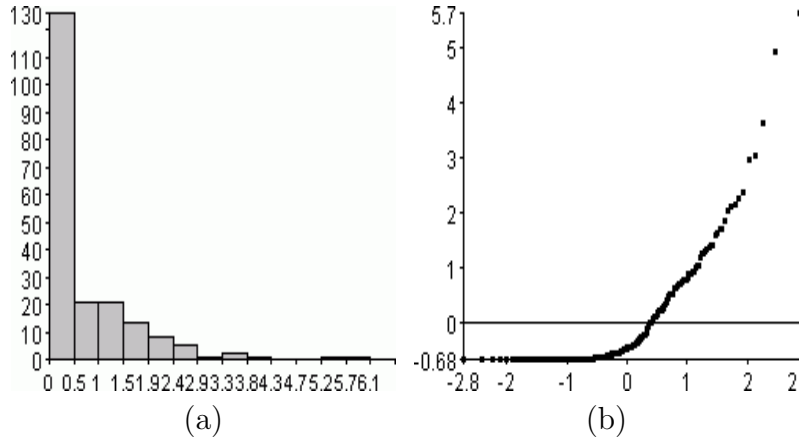



**Figure 4.1**: *Histogram and QQ plot of skewed trait distribution.*

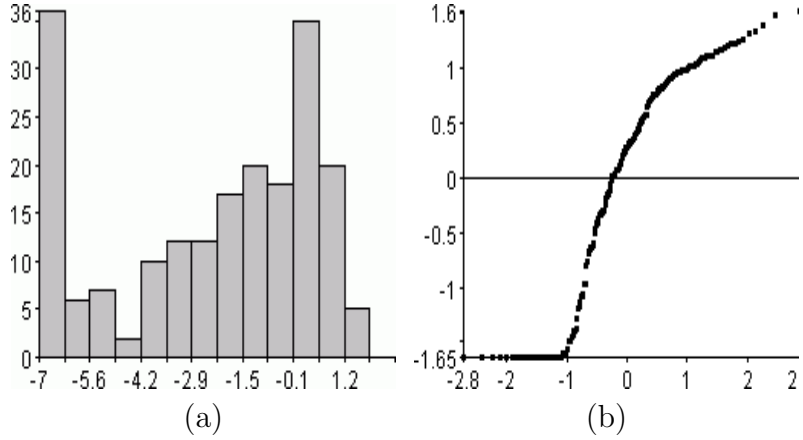
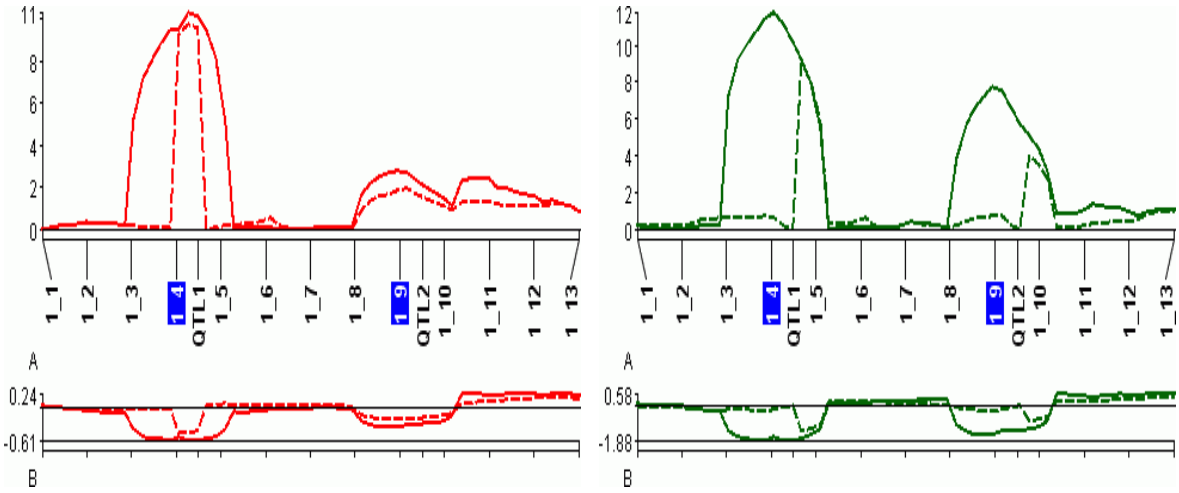



**Figure 4.2**: *Histogram and QQ plot of log-transformed skewed trait distribution.*

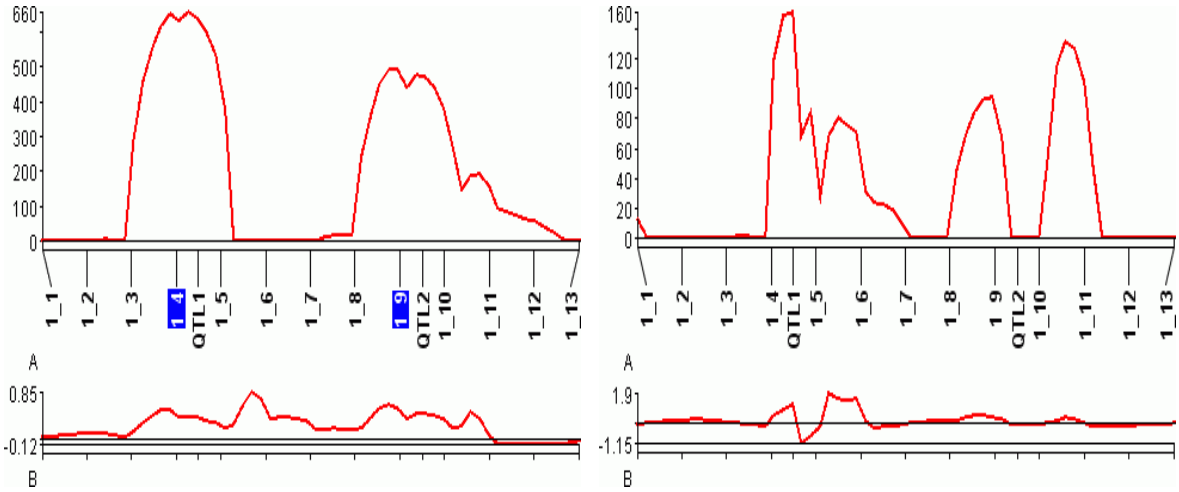
In this case, normal-based methods generally could not correctly detect $Q_2$, the QTL with the weaker effect. After log transformation, however, both CIM and MIM could detect both QTLs, as shown in Figure 4.3. The QTL effect estimates were wrong. For example, in Figure 4.3 (b), the estimate of the effect of $Q_1$ is -1.88. Its inverse transform is $\exp(-1.88) \approx 0.153$, which is far below the true value, 1.

GLZ-CIM could detect both QTLs in general, as shown in Figure 4.4 (a). However, the




(a) (b)

**Figure 4.3**: *Comparison of CIM and MIM on log-transformed and original skewed trait data. (a) CIM and MIM, which assume normality, could not detect QTL $Q_2$; (b) CIM and MIM could detect both QTLs. Solid line: CIM, broken line: MIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome. Markers selected as cofactors in CIM and GLZ-CIM are highlighted.*




(a) (b)

**Figure 4.4**: *Comparison of GLZ-based CIM and MIM on skewed trait data. The LOD and additive effect plots for (a) GLZ-CIM and (b) GLZ-MIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome. Markers selected as cofactors in CIM and GLZ-CIM are highlighted.*

QTL effect estimates are both incorrect. In Figure 4.4, the additive effect estimates for QTL $Q_1$ and $Q_2$ are 0.368 and 0.453.

Although in general GLZ-MIM could detect both QTLs according to the chosen detection criterion, it may have numerical stability problems, as illustrated in Figure 4.4 (b). This problem may arise partly because GLZ-based MIM is dependent on cofactor selection methods that assume an F distribution in determining the threshold at which a cofactor is selected or dropped. This threshold can be converted to a LOD threshold by a method specified in (Doerge 1995). The F-based threshold is used in regression covariate selection in standard statistical software, such as SAS and Minitab. But this threshold assumes normality, which does not hold here. The conversion to LOD score was performed because all LOD scores asymptotically approach a scaled noncentral $\chi^2$ distribution, regardless of normality. However, the LOD-based threshold appears to work poorly in this scenario.

### 4.8.2 Discretization case

For the discretization case, all methods could detect all QTLs correctly in all datasets. The QTL effect estimates in normal-based CIM and MIM were incorrect. This was expected because although the ordinal values can still be treated as numeric values, they are not exactly the same. QTL effect estimates are not available in GLZ-based methods.

### 4.8.3 Multicategorization case

For the multicategorization case, none of the normal-based methods could detect any of the QTLs, while GLZ-CIM and GLZ-MIM could detect all of them (Figure 4.5). In many cases, the peaks given by GLZ-MIM were lower than those of GLZ-CIM (Figure 4.5 (b)). However, both peaks passed the significance threshold obtained from permutation analysis.

### 4.8.4 XOR logic case

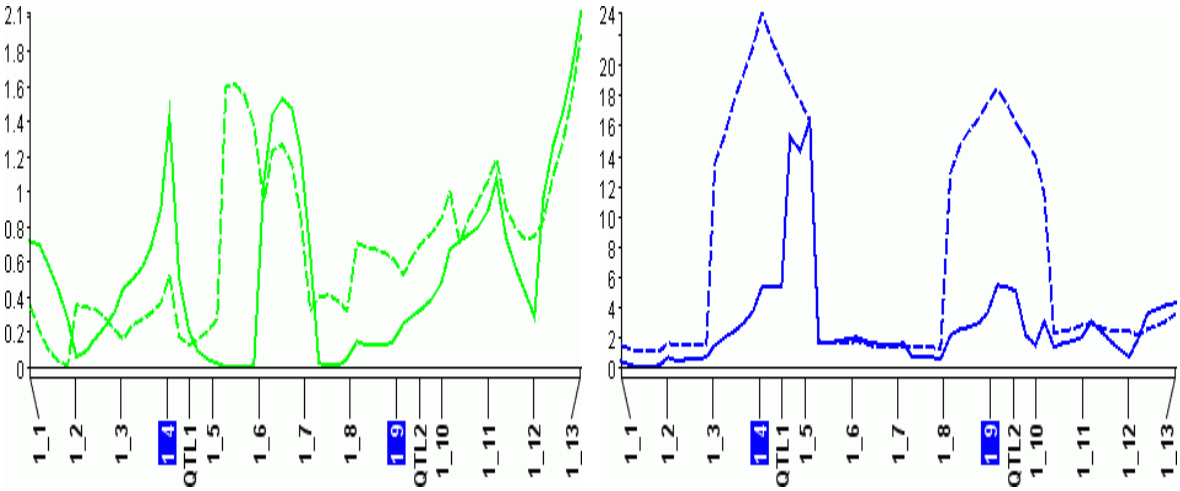
For the XOR logic case, none of the methods could detect any QTLs.

## 4.9 Discussion and conclusion

For traits with highly skewed distributions, GLZ-based methods were superior to normal-based methods for detecting QTLs. No methods could estimate the QTL effects correctly. Although normal-based methods could detect all QTLs after a log transform had been performed on the trait data, the estimated QTL effects were far from the correct values.

For ordinal traits, methods based on the assumption of normality were shown to be effective at detecting QTLs. In these data types, estimation of QTL effects is no longer accurate because parts of the general means may be incorporated into QTL effects. The GLZ-based method could be used to estimate QTL effects more correctly in numerical data. However, for ordinal data, effect estimation is not available under these methods.

For multiple-categorical data, GLZ-based methods are superior to methods based on normality. In most cases, GLZ-CIM is faster than GLZ-MIM, which took a long time to select QTL cofactors. The resulting LOD profile was also not as clear as that of GLZ-CIM, a problem perhaps related to the cofactor selection problem.




(a) (b)

**Figure 4.5**: *(a) CIM and MIM, which assume normality, detected no QTLs; (b) GLZ-based CIM and MIM detected all QTLs. Solid line: GLZ-MIM, broken line: GLZ-CIM. Vertical axis represents LOD score. Horizontal axis represents genetic location on the chromosome. Markers selected as cofactors in CIM and GLZ-CIM are highlighted.*

For non-linearly-separable data, such as in the XOR logic case, GLZ solutions that are based on probit or logit links may not be appropriate. If a more appropriate choice of link function and distribution are used, this problem may be soluble. For example, an inverse link function $a = b^{-1}$, where $a$ and $b$ are the encoded QTL genotypes, may be appropriate. Let $t_1 < t_2$ be threshold values. The XOR logic problem can be solved by identification of $t_1$ and $t_2$ values such that the region $ab < t_1$ contains all zeroes and the region $ab > t_2$ contains all ones. However, the binomial distribution will not work with this link function, which does not produce values between 0 and 1 corresponding to the binomial parameter $p$. Thus, the choice of link-function–error-distribution combination is still unclear. Until this problem is resolved, methods based on machine learning, such as decision trees and artificial neural networks, may be used.

# Chapter 5

# Clustered eQTL analysis with graphical Gaussian models

## Abstract

Several studies in expression QTL (*eQTL*), or QTL analysis applied to expression traits (*e-traits*), have been performed in various species for reconstructing gene regulatory networks, or diagrams representing a system of interrelationships among products of genes. Most of such studies describe extensions of available methods for reconstructing such networks from expression data, and do not incorporate genotype data. One such method, graphical Gaussian modeling (GGM), which relies on linear shrinkage of the complete partial correlation matrix among gene expression values, has been claimed to be effective for such network reconstruction. I develop an extension of GGM that incorporates eQTL information (GGM-eQTL). However, naïve application of GGM yields a network that appears too complex for meaningful biological inference, even after the incorporation of eQTL information. For this reason, a clustering method is applied for localizing gene network reconstruction. Because genes may have multiple biological functions, a clustering method (Fuzzy-K) accommodating multiple cluster memberships per gene is chosen. Because Fuzzy-K relies on visual inspection instead of objective methods for determining cluster membership cutoff values, a bootstrap method for empirically determining these values is developed. This Fuzzy-K plus bootstrap produces biologically plausible clusters of e-traits based upon their functional

annotations. GGM-eQTL is used to reconstruct the gene regulatory subnetwork putatively represented by each cluster. Since clusters overlap, some edges are detected in multiple clusters. These edges may be used to join subnetworks.

## 5.1 Introduction to eQTL analysis

EQTL analysis, or "genetical genomics", is QTL analysis applied to expression traits, or *e-trait*s (Jansen and Nap 2001). It has been performed in several studies, such as in yeast (Brem *et al.* 2002), human (Schadt *et al.* 2003; Watts *et al.* 2002; Steinmetz *et al.* 2002), and *Arabidopsis* (West *et al.* 2007). Although in most studies "e-trait" refers to mRNA expression obtained from microarray hybridization for each progeny in the mapping population, it may also refer to metabolite or protein levels. Since each e-trait is associated with an expressed DNA sequence, or *gene*, it has a physical location, expressed in bases or megabases (Mb) from one end of the physical map of a chromosome.

EQTL analysis aims to estimate the location and the effect of expression QTLs, or eQTLs, that account for the genetic variation responsible for the e-traits. This may reduce the error in the estimates from subsequent analyses, such as in construction of *regulatory networks*, or diagrams representing a system of interrelationships among products of genes. Figure 5.1 shows a typical genome-wide LOD-score plot for one e-trait. QTLs declared as significant are marked as black bars in the strip below. This process is repeated for all e-traits, ordered by their physical locations, to produce a *heat map* as shown in Figure 5.2. For methods such as composite interval mapping (CIM) that lack precision, an exclusionary window, or a minimum distance among the declared eQTLs, may be enforced to reduce false eQTL detection (West *et al.* 2007).

The heat map constructed from the eQTL analysis shows the eQTLs whose locations coincide with the physical locations of a given e-trait (*cis*-eQTLs) and also those that do not (*trans*-eQTLs). The physical location of an e-trait can be translated to or interpolated from a genetic location expressed in cM as described in Chapter 1. When an eQTL is

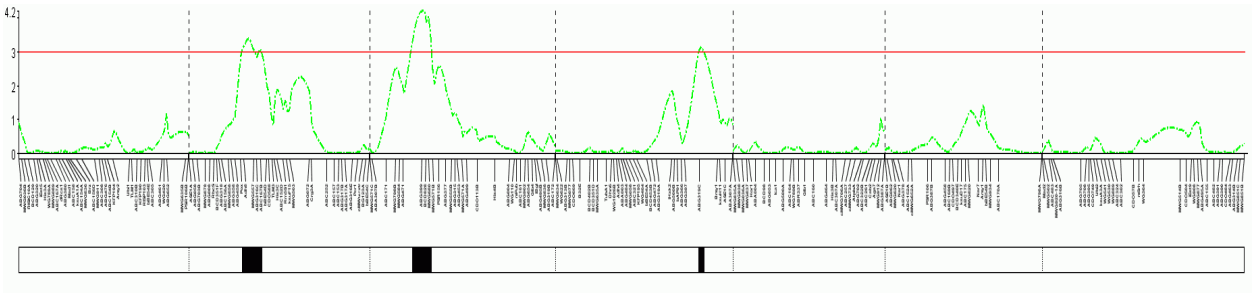


**Figure 5.1**: *Conversion of a QTL LOD plot into a heat map for one e-trait.*

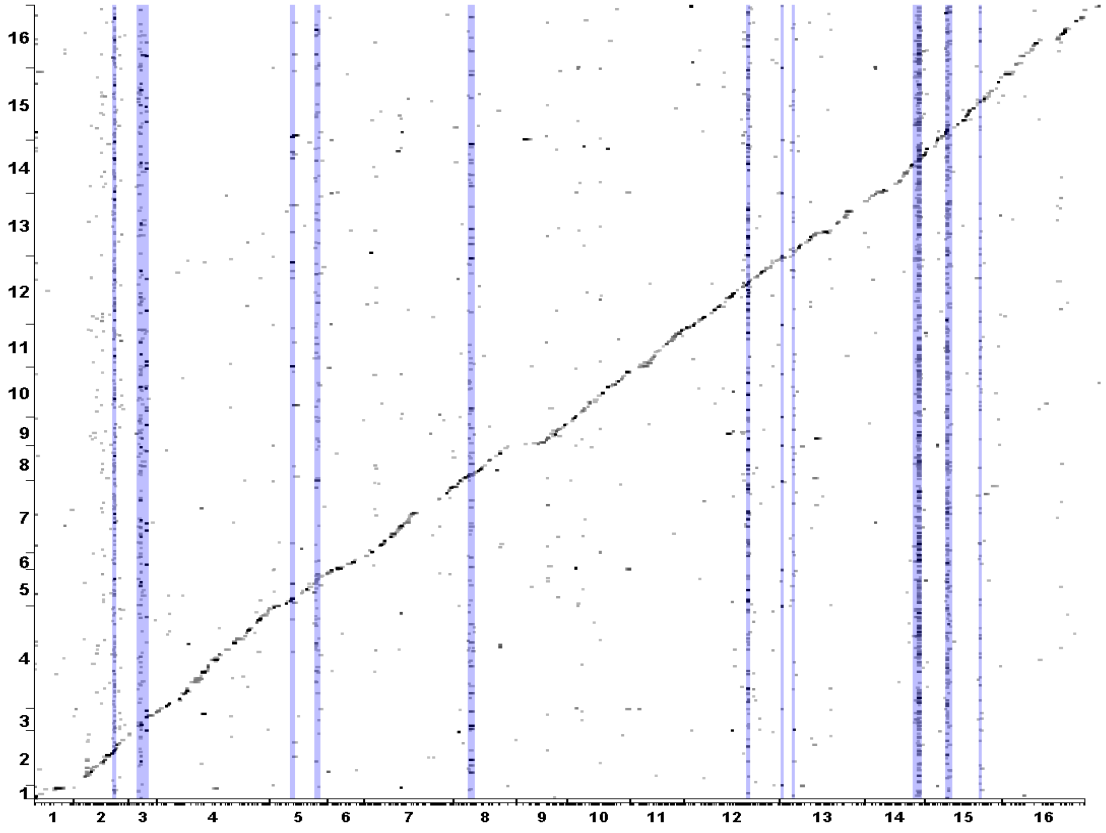



**Figure 5.2**: *EQTL heat map in QGene. The heat map was produced from MIM analysis of (*Brem *et al. 2002) data, with hotspots (see text) highlighted.*

within a certain distance of the corresponding genetic location of an e-trait, it is declared as a *cis*-eQTL (West *et al.* 2007). Since both physical and genetic locations are ordered by magnitude along their axes, the *cis*-eQTLs will form a diagonal pattern, while *trans*-eQTLs form vertical patterns, as shown in Figure 5.2.

The distribution of eQTL genetic contributions, or $R^2$, in *cis*- and *trans*-eQTLs suggests that the transcript levels of most genes are controlled by multiple eQTLs. The majority (89%) of eQTLs have $R^2 \leq 0.2$ (West *et al.* 2007). The contribution of *cis*-eQTLs to e-trait variation is known to be more than that of *trans*-eQTLs (Keurentjes *et al.* 2007).

Some of the unexplained variation is due to *transgressive segregation*, where the e-trait distribution for the offspring is wider than that of the parents. In transgressive segregation, the e-trait distributions of the two alleles overlap, making it harder to detect significant difference between the allele effects (Brem *et al.* 2002).

The genetic location associated with eQTLs above a certain threshold is called a *hotspot*. The significance threshold above which a region is declared a hotspot is estimated from a permutation procedure. It was proposed that hotspots indicate transcription factors (Jansen and Nap 2001), but this is not always the case (Yvert *et al.* 2003; Breitling *et al.* 2008).

## 5.2 Review of methods for pathways reconstruction from eQTL experiments

It was proposed that metabolic or chemical pathways can be reconstructed with the use of *cis*- and *trans*-eQTL data (Jansen and Nap 2001). The existence of common eQTLs in a set of e-traits was assumed to imply that the e-traits are directly related. Let $\mathbf{Q}_i$ represent the set of eQTLs for e-trait $i$ and $\subset$ the subset relationship. The authors proposed that if $\mathbf{Q}_A \subset \mathbf{Q}_B \subset \mathbf{Q}_C$, it may be inferred that gene $A$ "influences" gene $B$, which in turn influences gene $C$. However, because of set-theory considerations this method may be overly simplistic, and it does not seem to have been applied in practice.

A *graph*, or a system of *nodes* and *edges* that represents interconnections between objects,

is often used as the internal representation of a pathway. The nodes represent e-traits or eQTLs, while the edges (or links) represent the relationship between a pair of nodes. The graph constructed by these methods may be *directed* or *undirected*, depending on the type of the edges. In directed graphs, the notation $A \rightarrow B$ denotes a directed edge from node $A$ to node $B$, which means that variation in the object represented by node $A$ is the cause of that represented by node $B$. In undirected graphs, the notation A—B denotes an undirected edge, which means nodes $A$ and $B$ are related with no information of causal relationship.

The more common use of network inference is on microarray data acquired for clones of identical genotypes across multiple conditions and time points. In contrast, expression variation in an eQTL experiment is across individuals of different genotypes, rather than across environmental conditions. In principle, this distinction means that eQTL networks can be given directed edges corresponding to genetic causation.

In the following subsections, some studies in pathway reconstruction with eQTL information are reviewed. In all of these studies the e-traits are derived from mRNA expression data.

### 5.2.1 Bayesian network

A Bayesian network (BN) (Pearl 1988) is a probabilistic directed acyclic graphical model that represents a set of random variables and their probabilistic dependencies. The network is a compact representation of the joint probability of the random variables. A BN learning algorithm (Heckerman 1998; Neapolitan 2003) is an algorithm for reconstructing such a network from empirical data. These algorithms have also been used to reconstruct gene regulatory networks from expression data in studies such as (Friedman *et al.* 2000).

Learning algorithms to find the optimal network are mostly computationally intractable ("NP-hard") (Chickering *et al.* 1994, 2004), with algorithmic complexity strictly greater than $O(n^{\log_k N} N)$, where $n$ is the number of nodes or variables, $N$ is the number of data

points, and $k$ is the average number of possible values of each node (Dojer 2006). In practice, however, $N$ grows exponentially as the number of false edges decreases.

Most popular learning algorithms use some heuristics to find the optimum network, such as node ordering (Cooper and Herskovitz 1992), ignoring global dependencies when computing posterior probability (Spiegelhalter *et al.* 1993), and greedy searching (Heckerman 1998). Though these algorithms may not find the optimum network for the data, they work well in practice. The algorithm proposed by Friedman *et al.* (2000) focuses on local structures (the subnetwork that contains the currently examined node and nodes "close" to it) and uses greedy search to add and remove nodes and edges.

In addition to the computational issues, BN learning algorithms usually require more data than is common in current microarray experiments. With less data, the algorithm may output a different network that explains the data as well as the optimum network (Bing and Hoeschele 2005). This is also known as an *aliasing* problem.

Zhu *et al.* (2004) imposed two simplifying assumptions to address the issues of computation and aliasing. First, a gene cannot be controlled by more than three genes. Second, the possible parents or controllers for a given gene are limited to a subset of genes, instead of the complete gene set. The selection of candidate parents for a particular gene for each individual chromosome is based on weighted average correlations, given by:

$$r_{xy} = \sum_c w_{xy}(c) r_{xy}(c), \qquad r_{xy}(c) = \frac{\sum_l x_c(l) y_c(l) I_{c_{xy}}(l)}{\sum_l x_c(l)^2 + y_c(l)^2}$$

$$w_{xy}(c) = \max_l [\min(x_c(l), 10) \times \min(y_c(l), 10)]$$

where $x_c(l)$ and $y_c(l)$ are LOD scores at locus $l$ on chromosome $c$ and $I_{c_{xy}}(l)$ is defined as:

$$I_{c_{xy}}(l) = \begin{cases} 1, & \text{if } x_c(l) > 1.5, y_c(l) > 1.5 \\ 0, & \text{otherwise} \end{cases}$$

The $80^{\text{th}}$ percentile of the rank-ordered list of correlations is arbitrarily chosen as the cutoff.

To incorporate eQTL information and to differentiate eQTL co-location for a pair of e-traits due to pleiotropy from multiple closely linked eQTL, Zhu *et al.* (2004) calculated a

mutual information measure for each pair of e-traits:

$$\mathrm{MI}(A, B) \sum_{i,j} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)}$$

The $80^{\text{th}}$ percentile of the rank-ordered list of this score was arbitrarily chosen as the cutoff. The intersection of the sets obtained from the correlation and mutual information scores were the set of candidate parents for a particular gene.

Since correlation and mutual information scores are symmetric, they can indicate only association and not causality. For determining the causal relationship between genes $X$ and $Y$, Zhu *et al.* (2004) defined the following metric:

$$p(X \to Y) = r_{xy} \frac{N(Y)}{N(X) + N(Y)}$$

where $N(X)$ is the number of significant eQTL for gene $X$. If $p(X \to Y) > p(Y \to X)$, we infer that the influence of gene $X$ on gene $Y$ is stronger than the reverse.

### 5.2.2 Partial correlation

Partial correlation can be used to reconstruct a gene regulatory pathway (Bing and Hoeschele 2005). This method implicitly assumes that e-traits that are controlled by a common set of eQTLs may be controlled by a common regulatory element. After eQTLs for each e-trait have been located and filtered by a false-discovery-rate (FDR) method, a confidence interval (CI) for each eQTL is determined with a bootstrap method. The CI is further refined by selective bootstrap sampling. Because according to the authors, "unnecessarily large QTL CIs can result from the presence of multiple QTL in the same chromosome affecting the same expression profile," a sliding three-marker regression is performed, where a marker $i$ is fitted along with its flanking markers $(i-1, i+1)$. If there is a single eQTL in the CI, only the two markers flanking the eQTL will have a nonzero expected partial regression coefficient. If there are two eQTLs in the CI, all three markers will have nonzero expected partial regression coefficients. In this case, the CI of each of the multiply linked eQTLs is defined by its flanking markers.

After the CIs of all eQTLs are determined, partial correlation coefficients are computed to construct the network. For each eQTL, the set of e-traits that lies within the CI is determined. The Spearman correlation coefficient is computed for each e-trait controlled by one eQTL and is tested for significant departure from zero via $t$-test. The P-values are then adjusted by the Bonferroni method. The e-trait, say $G_1$, with the most significant correlation coefficient is identified. Then first-order partial correlation coefficients, conditioned on $G_1$, are computed for the rest of the e-trait within the CI of the same eQTL. These coefficients are also tested similarly. The gene with the next highest correlation coefficient is identified. This process is repeated to identify the subsequent genes, with the partial correlation coefficients conditioned on the e-traits declared significant in the previous iteration. The partial correlation coefficients determine where the genes will be placed in the network and the direction of the edges. A network of genes is then constructed by the linking of these edges. Since a *trans*-eQTL affects many different e-traits, the gene network can be connected across all e-traits. Gene Ontology (GO) (Gene Ontology Consortium 2000) is then used to annotate the genes and identify subnetworks.

### 5.2.3 Ranking

In this method (Keurentjes *et al.* 2006, 2007), the e-traits whose locations are within the confidence interval of each identified eQTL are grouped according to functions specified in the GO and literature search since e-traits of the same functional group, *e.g.,* flowering genes, are likely to be coregulated. The grouping is intended to find the "master regulator". The e-trait that best correlates to the other e-traits in the group is chosen as a candidate. The e-traits in a functional group are ranked by the iGA method (Breitling *et al.* 2004) to identify the subgroups. Then the Spearman correlation is computed for each e-trait pair within each subgroup. The regulator of each subgroup is defined as the e-trait with the highest correlation score with other e-traits in the subgroup. Partial correlation coefficients

are computed to determine the edges between e-trait pairs, starting with the regulator e-trait, as in Bing and Hoeschele (2005).

### 5.2.4 Clique

This method (Chesler *et al.* 2005; Baldwin *et al.* 2005) attempts to find *cliques*, or sets of completely interconnected nodes, in a graph. After the eQTLs for each e-trait are identified by composite interval mapping (CIM) (Zeng 1994), correlation scores are computed for the e-traits whose locations lie within the CI of a given eQTL. A high-pass filter is applied to the correlation scores and only edges with $|r| > 0.85$ are preserved. Since correlation scores are symmetric, the resulting graph is undirected. Cliques are then searched with a specialized vertex-cover algorithm from the undirected graph from the filtered correlation score (Baldwin *et al.* 2005). Since the vertex-cover problem (the problem of finding the smallest set of nodes such that each edge in the graph is connected to at least one node of the set) is "NP-complete", a so-called "crown reduction" algorithm (Baldwin *et al.* 2005) is performed to reduce the search space.

### 5.2.5 Likelihood-based model selection

This method uses a likelihood ratio test to select models among eQTLs (L), e-traits (R), and complex traits (T) (Schadt *et al.* 2005). The three models tested in the method are: $M_1$) $L \rightarrow R \rightarrow C$, $M_2$) $L \rightarrow C \rightarrow R$, and $M_3$) $L \rightarrow R$, and $L \rightarrow C$. $M_1$ represents a causal relationship, where an allelic difference in the eQTL $L$ acts on $C$ through transcript $R$. $M_2$ represents a reactive relationship, where $R$ is modulated by $C$. $M_3$ represents an independent relationship, where $L$ acts on $R$ and $C$ independently.

The eQTLs for the models are detected by CIM with cofactors selected by forward stepwise regression. The cofactors are limited to a maximum of six markers. It is unclear whether Schadt *et al.* (2005) applied an FDR adjustment, but this is implied. Multiply linked and pleiotropic eQTL are tested by an adjusted method of multiple-trait QTL analysis (Jiang and Zeng 1995).

Models $M_1$, $M_2$, and $M_3$ are tested by likelihood equations that describe the conditional dependencies among the eQTL, the transcript, and the conventional quantitative traits. The tests are as follows:

$$M_1.\ P(L,R,C) = P(L)P(R|L)P(C|R)$$
$$M_2.\ P(L,R,C) = P(L)P(C|L)P(R|C)$$
$$M_3.\ P(L,R,C) = P(L)P(R|L)P(C|R,L).$$

The relationship of $L \rightarrow C$ holds if $L$ is an eQTL of trait $C$. The direction of the relationship between $R$ and $C$ is determined from the mutual information measure. The Akaike information criterion (AIC) score is computed for these models and the model with the highest score is chosen. The rest of the e-trait gene network is constructed by the BN learning algorithm similar to that of Zhu *et al.* (2004). The generalized linear model is used to assess the degree of epistatic interactions among all pairwise positions.

### 5.2.6 Module

This method (Lum *et al.* 2006) is an extension of the method described by Schadt *et al.* (2005). The e-traits are grouped into several "modules" by adjusted Pearson correlation score and hierarchical clustering. An iterative algorithm is used to maximize a "module efficiency" score. The relationship between conventional trait and e-traits is established exactly as described in Schadt *et al.* (2005). The rest of the gene network is constructed with the BN learning algorithm proposed by Zhu *et al.* (2004).

### 5.2.7 Markov score

In this method (Kulp and Jagalur 2006), *cis*-eQTLs are detected by an interval mapping method, while *trans*-eQTLs are detected by the same method but with coregulated e-traits as fixed-effect covariates. This method was claimed to detect weaker *trans*-eQTL better than existing methods. The scores for determining significance are derived from the Markov

score, which is a log of posterior odds. In addition to *cis*- and *trans*-eQTLs, this method can detect a third class of eQTLs, called *cis-trans*-eQTLs, where an eQTL influences a transcript through another transcript.

### 5.2.8 SEM

Methods based on structural equation modeling (SEM) have been proposed (Liu *et al.* 2008; Li *et al.* 2006) to overcome the limitation of acyclicity imposed by the Bayesian network (BN). Although dynamic BN (DBN) (Murphy and Mian 1999) allows such cyclic structure for discrete time points, it requires time-series data, which has not been produced by eQTL studies to date. Thus, the SEM method is used to search for the structure instead, with estimates computed by a maximum likelihood method.

#### Principal components eQTL

In (Liu *et al.* 2008), $k$-means with absolute correlation as the distance measure was used to cluster the e-traits into 100 clusters. Although multiple-trait analysis provides more power for detecting pleiotropic QTL (Jiang and Zeng 1995), the eQTL analysis was performed on principal components (PC) instead. PC-based QTL analysis was claimed by Mähler *et al.* (2002) to work well in practice. The eQTLs found from analyzing the PCs were called PC-eQTLs and assumed to be common regulators of all e-traits with high loadings. The *cis*- and PC-eQTLs were identified by single-marker analysis with a Kruskal-Wallis test, while *trans*-eQTLs were identified by a method similar to the Markov score method (Kulp and Jagalur 2006). Instead of conventional interval mapping method, a regression-based method and the intersection–union test (IUT) were used. Candidate regulators whose closest marker had a recombination fraction of at most 0.25 with the marker closest to the target e-trait were excluded to avoid collinearity problems. An "encompassing directed network" was then constructed by assignment of directed edges from eQTL to *cis*-regulated e-trait nodes, from *cis*-regulated to *cis-trans*-regulated e-trait nodes, from *trans*-regulator to other e-trait nodes, and from *trans*-eQTL to other e-trait nodes.

SEM is computationally intensive. Most SEM algorithms can be applied to a system with only tens of variables. The authors managed to speed up the algorithm to allow hundreds of variables.

**Path refinement**

In (Li *et al.* 2006), the eQTLs were detected by single-marker regression. Pleiotropic eQTLs were detected by a multiple-regression model with other e-traits as fixed-effect covariates. After eQTLs were found, an initial path model of the SEM was constructed with eQTLs as root nodes. The model was then assessed by a goodness-of-fit test that follows a $\chi^2$ distribution. In subsequent iterations, the model was refined by addition, deletion, or reversal of edges.

## 5.3 Graphical Gaussian modeling

### 5.3.1 Introduction to GGM

When the number of genes $p$ greatly exceeds the number of observations $n$, the unbiased and maximum-likelihood estimate of variance–covariance matrix $\mathbf{S}$ of the expression traits (mathematically related to the correlation matrix used for network inference) will not be a good estimate of the true covariance matrix $\mathbf{\Sigma}$ since the eigenvalues of $\mathbf{S}$ differ greatly from the eigenvalues of $\mathbf{\Sigma}$. However, many expression data methods that rely on the correct estimate of $\mathbf{\Sigma}$ use $\mathbf{S}$ instead. This may lead to accuracy problems, especially since $\mathbf{S}$ is likely to be almost singular.

A shrinkage method called graphical Gaussian modeling (GGM) was proposed by Schäffer and Strimmer (2005) to overcome the eigenvalue divergence in variance–covariance estimation. Let $\mathbf{S}^*$ be a linear shrinkage of the variance–covariance matrix, $\mathbf{T}$ the shrinkage target, and $\lambda$ the shrinkage "intensity". $\mathbf{S}^*$ is defined as

$$\mathbf{S}^* = \lambda\mathbf{T} + (1 - \lambda)\mathbf{S}$$

Although there are many shrinkage targets to choose from, $\mathbf{T} = \text{Diag}(\mathbf{S})$ is chosen as a good compromise between the required number of estimated parameters and the full specification of covariance or correlation model. Let $\mathbf{Y} = [y_{ki}]$ be the original data matrix of $n \times p$, $\bar{y}_{\cdot i}$ the mean of the $i^{\text{th}}$ column of $\mathbf{Y}$, $w_{kij} = (y_{ki} - \bar{y}_{\cdot i})(y_{kj} - \bar{y}_{\cdot j})$, and $\bar{w}_{\cdot ij} = n^{-1} \sum_{k=1}^{n} w_{kij}$. The corresponding shrinkage intensity is $\lambda = \min(0, \max(1, \lambda^*))$, where

$$\lambda^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}, \quad \text{and } \widehat{\text{Var}}(s_{ij}) = \frac{n}{(n-1)^3} \sum_{k=1}^{n} (w_{kij} - \bar{w}_{\cdot ij})^2, \quad s_{ij} = \frac{n}{n-1} \bar{w}_{\cdot ij}$$

$\lambda$ is formulated to minimize the quadratic risk function, or equivalently the variance of the differences between the entries of the true variance–covariance matrix $\mathbf{\Sigma}$ and those of the standard estimate $\mathbf{S}$, *i.e.*, $\text{Var}[\mathbf{vec}(\mathbf{\Sigma} - \mathbf{vec}(\mathbf{S}))]$ (Schäffer and Strimmer 2005). Since the standard estimate of the variance–covariance matrix $\mathbf{S}$ is unbiased, any other estimate of $\mathbf{\Sigma}$ will introduce bias. For this reason, it is preferable to keep $\lambda$ as close to 0 as possible.

I have reconstructed the following derivation, as none was provided by Schäffer and Strimmer (2005). For speed in computation, singular-value decomposition (SVD) is used to decompose $\mathbf{Y}$ into $\mathbf{UDV}'$, where $\mathbf{D}$ is a diagonal matrix of the eigenvalues. Because $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, $\mathbf{UU}' = \mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_n$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_p$. If $\mathbf{X}$ is standardized, the correlation matrix among e-traits is

$$\underset{p \times p}{\mathbf{R}} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} = \frac{1}{n-1} = \mathbf{VDU}'\mathbf{UDV}' = \frac{1}{n-1}\mathbf{VD}^2\mathbf{V}'$$

Since $\text{Diag}(\mathbf{R}) = \mathbf{I}_p$, the shrunken correlation matrix is

$$\begin{aligned} \mathbf{R}^* &= \lambda \mathbf{I}_p + (1-\lambda)\mathbf{R} \\ &= \lambda \left( \mathbf{I}_p + \frac{1-\lambda}{\lambda} \mathbf{R} \right) \\ &= \lambda \left( \mathbf{I}_p + \frac{1-\lambda}{\lambda(n-1)} \mathbf{VD}^2\mathbf{V}' \right) \end{aligned}$$

Since inverting $\mathbf{R}^*$ directly would be computationally expensive, the following matrix identity (Woodbury 1950) may be used:

$$(\mathbf{A} + \mathbf{UCT})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{TA}^{-1}\mathbf{U})^{-1}\mathbf{TA}^{-1}$$

Let $c = (1-\lambda)/(\lambda(n-1))$. Substitute $\mathbf{A}$ for $\mathbf{I}_p$, $\mathbf{C}$ for $\mathbf{I}_n$, $\mathbf{U}$ for $c\mathbf{VD}$, and $\mathbf{T}$ for $\mathbf{DV}'$ to yield

$$\begin{aligned}\mathbf{R}^{*-1} &= \frac{1}{\lambda}\left[\mathbf{I}_p - c\mathbf{I}_p\mathbf{VD}\left(\mathbf{I}_n + c\mathbf{DV}'\mathbf{I}_p\mathbf{VD}\right)^{-1}\mathbf{DV}'\mathbf{I}_p\right] \\ &= \frac{1}{\lambda}\left[\mathbf{I}_p - c\mathbf{VD}\left(\mathbf{I}_n + c\mathbf{D}^2\right)^{-1}\mathbf{DV}'\right] \quad (\text{Since } \mathbf{V}'\mathbf{V} = \mathbf{I}_n)\end{aligned}$$

Since $\mathbf{D}$ is a diagonal matrix, inverting $(\mathbf{I}_n + c\mathbf{D}^2)$ takes only linear time. Let $\mathbf{D}^* = c\mathbf{D}\left(\mathbf{I}_n + c\mathbf{D}^2\right)^{-1}\mathbf{D}$. Notice that $\mathbf{D}^*$ is also a diagonal matrix whose diagonal elements are

$$d^*_{ii} = \frac{cd^2_{ii}}{1 + cd^2_{ii}}$$

where $d_{ii}$ is the diagonal element of $\mathbf{D}$.

Let $\mathbf{\Omega} = \mathbf{R}^{*-1} = [\omega_{ij}]$. The shrunken partial-correlation matrix is given by $\mathbf{R}^{**} = [r^{**}_{ij}]$, where $r^{**}_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$. $r^{**}_{ij}$ is the partial correlation between genes $i$ and $j$ given all other genes, with linear shrinkage applied.

A $p$-value $p_{ij}$ for each entry of matrix $\mathbf{R}^{**}$ is computed based on a shrinkage-T distribution (Opgen-Rhein and Strimmer 2007). For each $p_{ij}$, an adjustment intended to minimize the false-positive or false-discovery rate (FDR), or the $q$-value (Storey and Tibshirani 2003), $q_{ij}$ is computed. If $q_{ij}$ is above a prespecified threshold, an undirected edge between gene $i$ and $j$ is drawn.

### 5.3.2 Extension of GGM for eQTL analysis

I propose an extension to the GGM method, called GGM-eQTL, that incorporates the eQTL information by conditioning the partial correlation matrix on eQTL genotypes. The purpose of the conditioning is to reduce errors in inferring the relationships between e-traits by improving the correlation scores between e-traits that share common eQTLs. Let $\mathbf{Q}_i$ be the set of eQTLs declared significant for e-trait $i$. For arbitrary sets $\mathbf{A}$ and $\mathbf{B}$, let $\mathbf{A} \cup \mathbf{B}$ denote their union, $\mathbf{A} \cap \mathbf{B}$ their intersection, $\mathbf{A}'$ the complement of set $\mathbf{A}$, $\mathbf{A} \setminus \mathbf{B} = \mathbf{A} \cap \mathbf{B}'$ the set difference, and $\mathbf{A} \ominus \mathbf{B} = (\mathbf{A} \cup \mathbf{B}) \setminus (\mathbf{A} \cap \mathbf{B})$ the (symmetric) difference between $\mathbf{A}$ and $\mathbf{B}$. For each e-trait pair $i$ and $j$, the partial correlation score $r_{ij}$ obtained in the

previous section is conditioned on $\mathbf{Q}_{i,j} = \mathbf{Q}_i \ominus \mathbf{Q}_j$, *i.e.*, holding constant the eQTLs not common to the e-traits. For e-traits lacking common eQTLs, $\mathbf{Q}_{i,j}$ will reduce to $\mathbf{Q}_i \cup \mathbf{Q}_j$. In this situation, holding constant all eQTLs controlling the e-traits will reduce the partial correlation score to the correlation between residuals of e-traits $i$ and $j$ given their respective eQTLs. This correlation between residuals is expected to equal zero. Common eQTLs are defined as eQTLs lying less than 15 cM apart.

The computation of the shrinkage factor conditioned on eQTL data $\widetilde{\lambda}$ can be performed by replacement of $w_{kij}$ in the previous section with $w_{kij} = (y_{ki} - \hat{y}_{\cdot i})(y_{kj} - \hat{y}_{\cdot j})$, where $\hat{y}_{\cdot i}$ is the fitted value of e-trait $i$ given $\mathbf{Q}_i$. Thus $w_{kij}$ is in a sense a cross-residual between e-traits $i$ and $j$. The remaining computation for $\widetilde{\lambda}$ is carried out just as in the original GGM method.

In the eQTL case, the matrix $\mathbf{R} = [r_{ij}]$ is replaced with the correlation matrix conditioned on eQTL data $\widetilde{\mathbf{R}} = [r_{ij}]$. Let $\hat{r}_{i,k}$ be the $k^{\text{th}}$ fitted value of e-trait $i$ on $\mathbf{Q}_{i,j}$. Let $\bar{\hat{r}}_i$ be the average of all $\hat{r}_{i,k}$. $r_{ij}$ can be obtained by

$$r_{ij} = \frac{\sum_{k=1}^{n} \hat{r}_{i,k}\hat{r}_{j,k} - n\bar{\hat{r}}_i\bar{\hat{r}}_j}{\sqrt{\sum_{k=1}^{n} \hat{r}_{i,k}^2 - \bar{\hat{r}}_i}\sqrt{\sum_{k=1}^{n} \hat{r}_{j,k}^2 - \bar{\hat{r}}_j}}$$

Matrix $\widetilde{\mathbf{R}}$ is then shrunk similarly:

$$\widetilde{\mathbf{R}}^* = \widetilde{\lambda}\mathbf{I}_p + (1 - \widetilde{\lambda})\widetilde{\mathbf{R}}$$

An inverse of $\widetilde{\mathbf{R}}^*$ is required for a partial correlation matrix conditioned on all e-traits. An SVD of $\widetilde{\mathbf{R}}^*$ is computed to reduce computational cost of direct inversion. Since $\widetilde{\mathbf{R}}^*$ is symmetric, the SVD reduces to eigenvalue decomposition, which yields $\widetilde{\mathbf{R}}^* = \widetilde{\mathbf{V}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}'$. Hence

$$\widetilde{\mathbf{R}}^* = \widetilde{\lambda}\left(\mathbf{I}_p + \frac{1 - \widetilde{\lambda}}{\widetilde{\lambda}}\widetilde{\mathbf{V}}\widetilde{\mathbf{D}}^2\widetilde{\mathbf{V}}'\right)$$

Let $\widetilde{\mathbf{D}}^* = \left(\frac{1}{\widetilde{\lambda}} - 1\right)\widetilde{\mathbf{D}}\left(\mathbf{I}_p + \left(\frac{1}{\widetilde{\lambda}} - 1\right)\widetilde{\mathbf{D}}^2\right)^{-1}\widetilde{\mathbf{D}}$. Notice that $\widetilde{\mathbf{D}}^*$ is also a diagonal matrix

whose diagonal elements are

$$\widetilde{d}^{*}_{ii} = \frac{\left(\frac{1}{\widetilde{\lambda}} - 1\right)\widetilde{d}^{2}_{ii}}{1 + \left(\frac{1}{\widetilde{\lambda}} - 1\right)\widetilde{d}^{2}_{ii}}$$

where $\widetilde{d}_{ii}$ is the diagonal element of $\widetilde{\mathbf{D}}$. Then

$$\widetilde{\mathbf{R}}^{*-1} = \frac{1}{\widetilde{\lambda}}\left[\mathbf{I}_p - \widetilde{\mathbf{V}}\widetilde{\mathbf{D}}^{*}\widetilde{\mathbf{V}}'\right]$$

The computation of the shrunken partial-correlation matrix, p-values, and FDRs are carried out just as in the original GGM method. The edges whose FDR values are above a certain threshold (0.05 was used here) are removed.

### 5.3.3 Results

Applying GGM to all 6,229 e-traits without conditioning on eQTL data yielded a graph that appears too complex for meaningful biological inference. Of a possible 19.4 million edges, more than 2 million were declared significant at threshold FDR = 0.05. The graph was too large to be rendered by Cytoscape (Shannon *et al.* 2003), a pathway-analysis software package. Removing edges of e-traits that did not share common eQTLs still left more than 250,000 edges. The resulting graph (Figure 5.3) is still too complex.

When the number of e-traits $p$ is very large, computing a $p \times p$ correlation matrix is time-consuming. Other matrix operations, such as matrix multiplications and additions, not only are time-consuming, but also require much computer memory. To reduce complexity and computation cost by reducing the dimensionality of the problem, I chose Fuzzy-K clustering (Gasch and Eisen 2002), discussed in the next section. I applied GGM-eQTL to each cluster, yielding a set of subnetworks. These will hereafter be called "localized" or "local" networks, in contrast to the "global" network calculated without clustering.

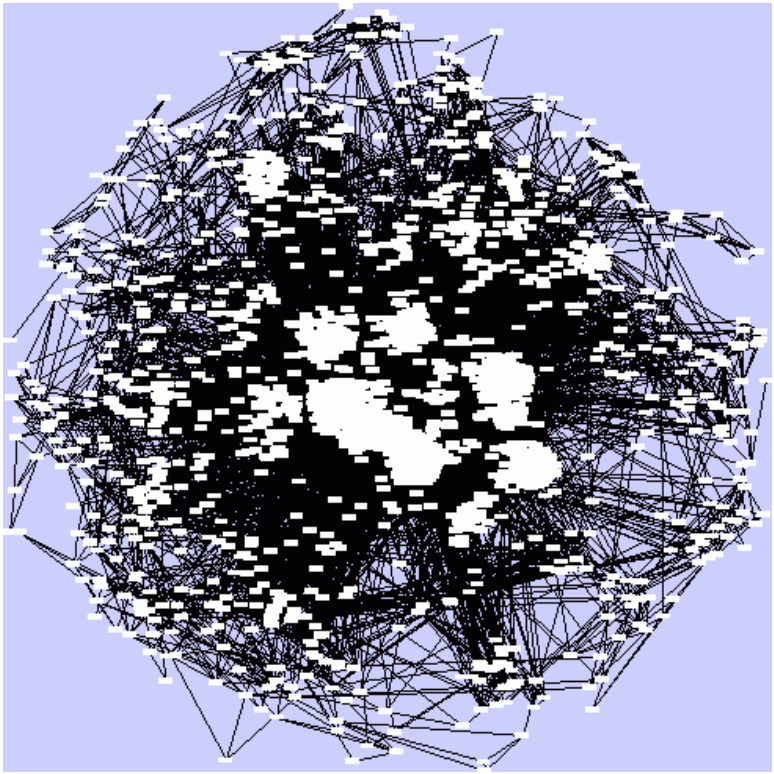


**Figure 5.3**: *A global gene regulatory network produced by GGM after removal of edges between e-traits lacking common eQTLs.*

## 5.4 Clustering in eQTL analysis

### 5.4.1 Introduction

Clustering has been used extensively for elucidating biological processes in microarray experiments. A survey of literature on this application is available (Jiang *et al.* 2004). One of the main uses of clustering is to aid in the inference of gene regulatory networks. The underlying assumption of the analysis is that genes of related function are often regulated under common control (Eisen *et al.* 1998). Clustering is used to uncover such control by identifying groups of similarly expressed genes. Clustering in some form was used in some eQTL analyses described in the preceding section. This was done to limit the search space for network construction or to identify similarities among groups of e-traits.

Because genes may exercise several, possibly distinct, biological functions, clustering methods that allow genes to belong to multiple clusters may be preferred. For example, ubiquinone has roles in respiration and oxidative stress in microbes (Søballe and Poole 1999). However, most clustering methods, such as the k-means algorithm, assign each object to only one cluster.

### 5.4.2 Fuzzy clustering

Fuzzy clustering describes a class of algorithms that allow objects to be classified into multiple clusters. Each object is associated with an array of real numbers, each of which reflects the degree of its membership in one cluster, or a *membership score*. For a given object, the sum of the membership scores across all clusters is 1.

Fuzzy k-means clustering (Bezdek 1981), an extension of k-means clustering, is an example of fuzzy clustering. In conventional fuzzy k-means, the number of clusters, $k$, is prespecified. The *centroids*, which are the vectors of the means of expression values in clusters, are initialized to random locations. Now a series of iterations begins in which first the membership scores of all $n$ objects for all $k$ clusters are computed according to a prespecified distance metric, and second, the centroids are updated with average coordinate vectors

weighted by membership scores. This process is repeated until convergence. Conventionally, a cutoff point is chosen such that a given cluster is defined as the set of genes whose membership scores for that cluster are above the threshold.

Fuzzy-K (Gasch and Eisen 2002) was developed as an extension of the fuzzy k-means algorithm. Since variants of k-means clustering are sensitive to the initial values of the centroids, principal components (PC) instead of random values are used as centroids. Pearson correlation is used as the distance metric. The clustering is done in $c$ rounds. Before the execution of the first round, Fuzzy-K has the option of weighting the genes and the observations based on the Pearson-correlation scores. Alternatively, the weights may be supplied in the data file. If no weights are supplied, Fuzzy-K will assume equal weights. In each round, genes are grouped into $k/c$ clusters using regular fuzzy k-means clustering. At the end of each round, all genes whose membership scores on a given cluster are at least 0.7 are accepted and set aside, while the remaining are reweighted and clustered in the next round. At the end of round $c$, all unaccepted genes are discarded. The centroids found in each round are collected into one set. Any two centroids that are highly correlated ($\rho \geq 0.9$) are replaced with a centroid that represents their average. Thus, the final number of centroids $k^*$ may be fewer than $k$. A membership matrix $\mathbf{M}$ is then constructed for all $n$ genes, yielding an $n \times k^*$ matrix. This matrix is then inspected with a visualization tool for selecting a membership cutoff for each cluster.

### 5.4.3 Fuzzy-K implementation notes

The original Fuzzy-K was implemented in C++. Some modification was required to make the source code supplied by the authors compilable and executable in the Linux operating system. The data type used in all computation processes was float (32-bit). The code contained undocumented changes, such as modification of the correlation threshold at the last round to 0.68, instead of 0.7.

I implemented Fuzzy-K in Java (JFuzzy-K). I used double (64-bit) as the data type of

the underlying computation in order to reduce numerical error due to repeated rounding. I used multithreading extensively in the clustering to speed up the computation.

The reimplementation was prompted by a wish to investigate the effect of the number of rounds on the number of genes successfully clustered. This modification could not be easily done since there were many hard-coded values and other bad coding practices in the original source code. In addition, the original implementation was inefficient. One run required about 4-5 hours, while the multithreaded reimplementation required less than 3 minutes.

### 5.4.4 Bootstrap method for fuzzy clustering

I developed a bootstrap method to overcome the limitation in Fuzzy-K of visual inspection to select a membership cutoff point. Let $m_{ij}$ be the membership score of gene $i$ in cluster $j$. Here, $m_{ij}$ is subject to the constraint $\sum_{j=1}^{k*} m_{ij} = 1$. Let $d_{ij}$ be the distance between gene $i$ and the centroid of cluster $j$. Write

$$m_{ij} = \frac{d_{ij}^{-f}}{\sum_{j=1}^{k^*} d_{ij}^{-f}}$$

where $f > 1$ is the "fuzziness factor". In most common implementations, including in Fuzzy-K, $f = 2$. The greater $f$ is, the more the clusters overlap.

For each cluster $j$, the null hypothesis for the bootstrap is $H_0 : \mathbf{m}_{\cdot j} = 0$, where $\mathbf{m}_{\cdot j}$ is the vector of membership scores of all genes on cluster $j$. At each bootstrap iteration, the centroid of cluster $j$ is shuffled (the numbers in the vector representing the centroid are scrambled), while the other centroids are held fixed. In each of the 1,000 iterations, the membership matrix is recomputed, $\mathbf{M}^*$ and the maximum value of $\mathbf{m}^*_{\cdot j}$ is recorded. The $\alpha$-level value of the $\max(\mathbf{m}^*_{\cdot j})$ values is then taken as the cutoff for cluster $j$.

Notice that in the computation of the cutoff of cluster $j$, $d_{ik}$ is constant in all iterations, for $j \neq k$. Hence, only $d_{ij}$ is recomputed and the $d_{ik}$s can be retrieved from the original distance matrix. This will speed up the bootstrap computation.

### 5.4.5 Experimental method

The bootstrap method for JFuzzy-K was performed on the data of Gasch and Eisen (2002) and Brem *et al.* (2002), with annotations retrieved from the Saccharomyces Genome Database (SGD) (Dwight *et al.* 2002). The first dataset was used for comparison with the clustering results of Gasch and Eisen (2002) and the second with those of Yvert *et al.* (2003).

Dataset G (Table 5.1) originally used by Gasch and Eisen (2002) to demonstrate Fuzzy-K comprises 6,153 genes and 93 environments from a variety of experiments imposing treatments such as zinc starvation, phosphate limitation, DNA-damaging agents, and various environmental stresses. The set of 4,373 genes whose standard deviation is greater than the criterion of $\log_2(0.45)$ selected by the authors is designated as dataset $G_{sub}$. The genes are arranged in rows and environments in columns. The expression values have been $\log_2$ normalized. Each dataset has its associated row and column weights.

**Table 5.1**: *Datasets used for Fuzzy-K bootstrap*

| Dataset | Source | No. genes | No. obs | Description |
|---|---|---|---|---|
| G | Gasch and Eisen (2002) | 6,153 | 91 | Original data |
| $G_{sub}$ | Gasch and Eisen (2002) | 4,373 | 91 | Subset of dataset A with $\sigma \geq \log_2(0.45)$ |
| B | Brem *et al.* (2002) | 6,229 | 112 | Original data |
| $B_{sub}$ | Brem *et al.* (2002) | 890 | 112 | Subset of dataset C with $\sigma \geq \log_2(0.45)$ |

The dataset used in eQTL analysis (Brem *et al.* 2002) comprises the genotypes of 112 haploid progeny from a cross of yeast strains BY and RM (dataset B, Table 5.1) and $\log_2$-normalized transcript levels of 6,229 genes, or e-traits. Since the yeast progeny were not subjected to diverse environmental conditions, the transcript variation across progeny is limited to genetic and random variation. Of 6,229 genes, only 890 show standard deviations greater than $\log_2(0.45)$ (dataset $B_{sub}$).

### 5.4.6 Bootstrap results

#### Bootstrap results for G and $G_{sub}$ data

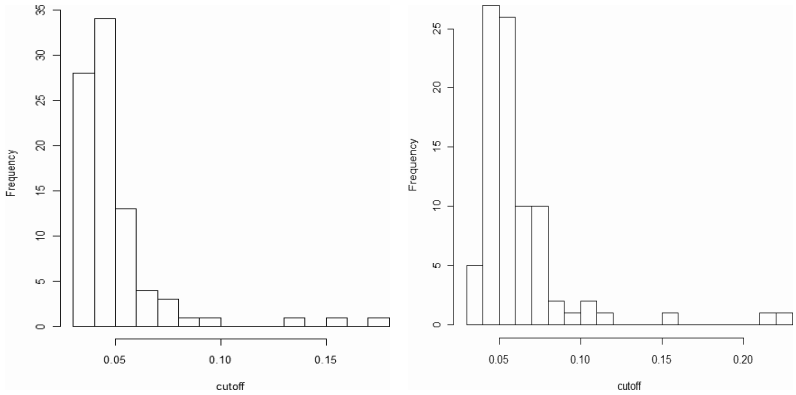
JFuzzy-K found 87 of 91 centroids reported by Gasch and Eisen (2002) for dataset $G_{sub}$, using the weights supplied in the file and setting $k = 120$ and $c = 3$ as suggested (Gasch and Eisen 2002). However, I could not replicate the calculations needed to estimate these weights. If the initial weights were reestimated with the method supplied in the code, only 72 centroids were found. If dataset G was used instead, only 78 and 61 centroids were found with and without reweighting, respectively. When the hacks and workarounds found in the original code were disabled, JFuzzy-K output one or two fewer centroids.

Although the clusters found by JFuzzy-K are not the same as those found by the original code, the alternative groupings are plausible based on the gene annotations. Of 87 centroids found by JFuzzy-K, only 5 are somewhat close to the reported centroids. The other centroids resemble linear combinations of the original centroids. For example, original cluster 40, which consisted of genes involved in nitrogen utilization (Gasch and Eisen 2002), is split into clusters 33 and 47. Transport and unknown genes in the original cluster 40 are replaced with other, mostly unknown, genes in the new clusters. The original cluster 68, which consisted of genes involved in glycolysis, is split into clusters 5, 9, and 21. The original cluster also contained genes for protein biosynthesis and degradation. The new cluster 21 groups the glycolysis genes, while clusters 5 and 9 contain the genes for protein biosynthesis and degradation, respectively. The original centroid 2 for the amino-acid biosynthesis cluster is about the same as centroid 1 from JFuzzy-K.

The bootstrap analysis also showed that the cutoff values varied from cluster to cluster. The authors did not provide the cluster cutoff values; only the list of genes of each cluster at cutoff values 0.06 and 0.08 was available as downloadable supplemental materials. At both cutoff values, the sets of genes at centroid 2 are identical with JFuzzy-K output, except for gene YMR300C encoding amidophosphoribosyltransferase. However, the bootstrap analysis

suggested a cutoff value of 0.033 at $\alpha = 0.05$. JFuzzy-K reported 8 more genes for that cluster.

The bootstrap results indicated that cutoff values of 0.06 and 0.08 correspond to $\alpha = 0.05$ and 0.01 for most clusters. These values were chosen by the original authors based on visual inspection. However, the histograms in Figure 5.4 contain some outliers, suggesting that visual inspection risks the spurious inclusion of genes in clusters.




(a) (b)

**Figure 5.4**: *The distribution of cluster cutoff values for data of* Gasch and Eisen *(2002) at $\alpha$ = a) 0.05, b) 0.01*

Increasing $k$ and $c$ always increased the number of detected centroids, contrary to the claim of Gasch and Eisen (2002). Values of $k = 622$ and $c = 10$ produced 234 unique clusters comprising 4,979 genes. Of these, 39 clusters contained 10 or fewer members, with only 3 clusters containing two or fewer. However, the more clusters found, the fuzzier they were. On average, a gene appeared in 5 different clusters, with some appearing in 15 clusters. If initial weights were estimated instead of equal weights assumed, only 195 clusters were found, covering 4,856 genes. The higher number of clusters identified at higher $k$ and $c$ values suggests a further subdivision of clusters detected at lower $k$ and $c$ values. At $k = 622$ and $c = 10$, many clusters could be combined into one that could be detected by the same algorithm at $k = 120$ and $c = 3$.

### Bootstrap results for B and $B_{sub}$ data

Since the size of BY×RM dataset is roughly the same as that of dataset G, the same parameter values were chosen for the clustering. At $k = 120$ and $c = 3$, 98 clusters were found, comprising 4,153 genes. Only 5 clusters had two or fewer members. On average one gene appeared in two clusters. If initial weights were estimated instead of equal weights assumed, only 85 clusters were found, covering 2,867 genes. Three of the clusters had two or fewer members. If only the 890 genes in dataset $B_{sub}$ were used to determine the centroids, 77 clusters covering 2,859 genes were found. In contrast, the previous hierarchical clustering (Yvert *et al.* 2003) found 205 clusters of more than two and 388 clusters of exactly two genes, covering 1,861 genes.

JFuzzy-K found clusters associated with known biological functions. For example, genes in cluster 8 (Table A.1) govern mostly sterol metabolism and transport. Clusters 5, 6, and 9 (Tables A.2–4) contain genes involved in amino-acid and protein biosynthesis. These clusters share some common genes.

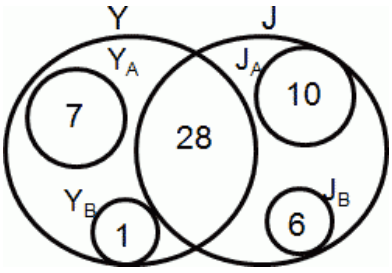
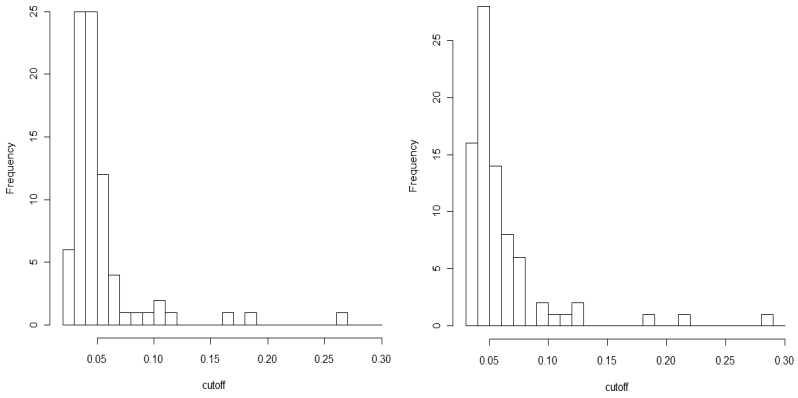



**Figure 5.5**: *Diagram showing the similarity of putative mating-associated clusters from JFuzzy-K and* Yvert *et al. (2003). Set Y: genes in cluster-group 5 of* Yvert *et al. (2003); J: genes in clusters 1, 7, and 39 from JFuzzy-K; $J_A$: other genes that may be involved in mating; $J_B$: genes that may be involved in other biological functions; $Y_A$: other mating genes not clustered by JFuzzy-K; $Y_B$: a "putative gene" placed in another cluster by JFuzzy-K. Numbers represent gene counts.*

A sample comparison of clusters produced by JFuzzy-K with one of the 13 cluster-groups into which Yvert *et al.* (2003) aggregated their 593 hierarchical clusters showed pronounced

overlap. 28 genes were shared between the 36 genes in cluster-group 5 and 44 genes in three JFuzzy-K clusters (Figure 5.5, Tables A.5–7).




(a) (b)

**Figure 5.6**: *The distribution of cluster cutoff values for the data of* Brem *et al. (2002) at* $\alpha$ = *a) 0.05, b) 0.01*

The overlap among clusters produced by JFuzzy-K suggests that some genes are involved in several related biological processes. For example, clusters 24 and 25 contain genes that may be involved in asexual reproduction (budding). These clusters share some genes with putative mating clusters.

The distribution of bootstrapped cluster cutoff values for the BY× RM cross data in Figure 5.6 was similar to that for the data of Gasch and Eisen (2002). The cutoff of a cluster containing only one gene of unknown function, YCL006C, is not shown since it is almost 1. The cutoff values of 0.06 and 0.08 also corresponded to $\alpha = 0.05$ and 0.01 for most clusters.

### 5.4.7 Discussion and conclusion

Comparisons suggest that JFuzzy-K clusters share the biological plausibility of the hierarchical and Fuzzy-K clusters identified by other authors. Where JFuzzy-K clusters do not match exactly those of Fuzzy-K, they represent plausible alternative groupings, based on the gene annotations.

The bootstrap extension provides objective estimates of cluster membership cutoffs, rendering visual inspection unnecessary. The distribution of the cutoff values as shown in Figures 5.4 and 5.6 suggests that a single cutoff for all clusters may not be appropriate. On the contrary, visual inspection to determine the cluster cutoff, aside from its impracticality, may be dangerous in view of the wide differences in cutoff values revealed by bootstrapping.

## 5.5 Localized GGM-eQTL

### 5.5.1 Methods

Fuzzy-K clustering with the bootstrap method as described in the preceding section was run for clustering the data from Brem *et al.* (2002) with parameters $k = 120$, $c = 3$, and bootstrap $\alpha = 0.05$. E-traits that were not clustered by the end of the iterations were discarded.

GGM-eQTL was applied to each cluster produced by JFuzzy-K, with the intent of building a subnetwork representing a biochemical pathway. To each cluster was applied an individual shrinkage intensity, $\lambda$, estimated as described in section 5.3.2. Since JFuzzy-K produced overlapping clusters, GGM-eQTL was expected to yield some edges shared by multiple clusters and thereby joining subnetworks. Cytoscape was used to visualize all networks.

The BioCyc Omics Viewer (Karp *et al.* 2002, 2005) was used to identify the yeast metabolic pathways associated with the 890 e-traits identified as varying in the eQTL dataset. Some of these pathways, with annotations retrieved from the BioCyc datase (Karp *et al.* 2005), were compared to the subnetworks produced by GGM-eQTL. The edges in the networks were also compared with the interactions recorded in the SGD. These represent interactions between proteins, derived from yeast-two-hybrid, affinity-precipitation, dosage-rescue, and "synthetic-interaction" experiments, along with some genetic interactions.

A bootstrap method was used to assess the reproducibility of the edges produced by GGM-eQTL. The bootstrap was repeated 1,000 times. The e-traits were arranged in rows

and yeast progeny in columns. The sampling for each iteration was done by permutation of the columns, preserving the correlation among e-traits. At each iteration, eQTL analysis, Fuzzy-K, and GGM-eQTL were applied to the permuted data. The frequency of detection of an edge in bootstrap iterations was recorded as a metric of its support. Edges in the original graph that appeared at least 50% of the iterations were retained.

### 5.5.2 Results

GGM-eQTL applied to clusters produced by Fuzzy-K clustering yielded a network connecting 1,766 e-traits as nodes with 20,167 edges (FDR=0.05). At $k = 120$, $c = 3$, and bootstrap $\alpha = 0.05$, Fuzzy-K found 77 clusters covering 2,859 e-traits, with 18 comprising five or fewer members. This means 1,093 e-traits clustered by Fuzzy-K were not connected at all. The subnetworks were constructed only on clusters with at least six members since FDR estimation requires "sufficient" data points. The shrinkage intensities ($\lambda$) were mostly between 0.1 and 0.2. Because the global $\lambda$ was 0.406, this implies that the localized correlation-matrix shrinkage process incurred less bias than the global one.

The clustering reduced the time needed to reconstruct networks. Without clustering, application of GGM to the data of Brem *et al.* (2002) required about 40 minutes, while GGM-eQTL required about two hours. With clustering, GGM e-QTL required only about 35 minutes.

Of the 890 sufficiently variable e-traits, only 215 appeared in BioCyc-curated metabolic pathways, as represented by red edges in Figure 5.7. The coverage of these 215 e-traits in the reference pathways is sparse. Only pathways with complete or partial coverage were chosen for evaluation of GGM-eQTL.

GGM-eQTL could construct putative gene regulatory networks from the variable e-traits, although they were not identical to the reference pathways. Figures 5.8–9 show putative amino-acid biosynthesis networks formed by GGM-eQTL from clusters 5, 6, and 9. The cluster shrinkage estimates are 0.103, 0.119, and 0.102. Each of these subnetworks includes

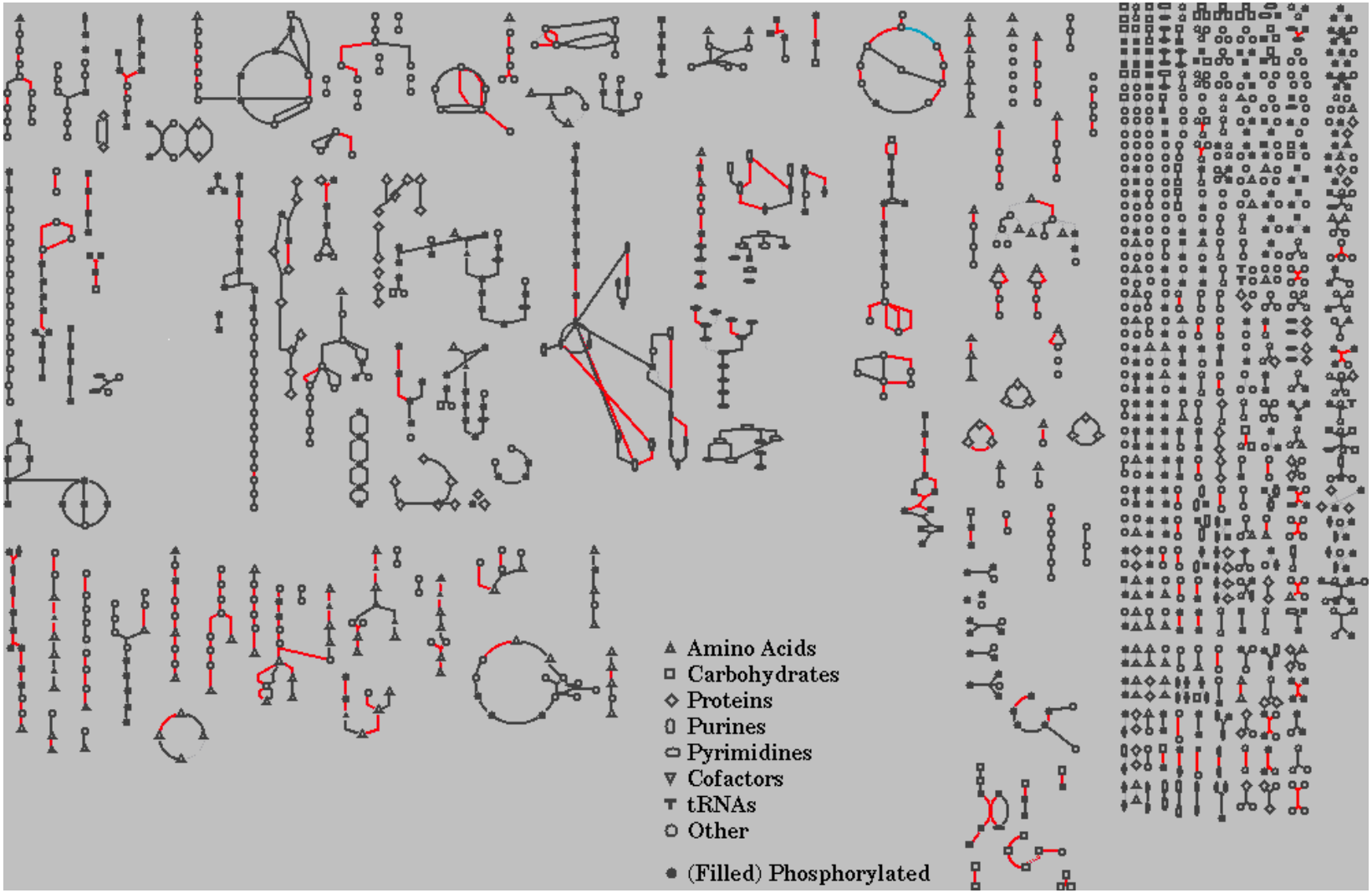
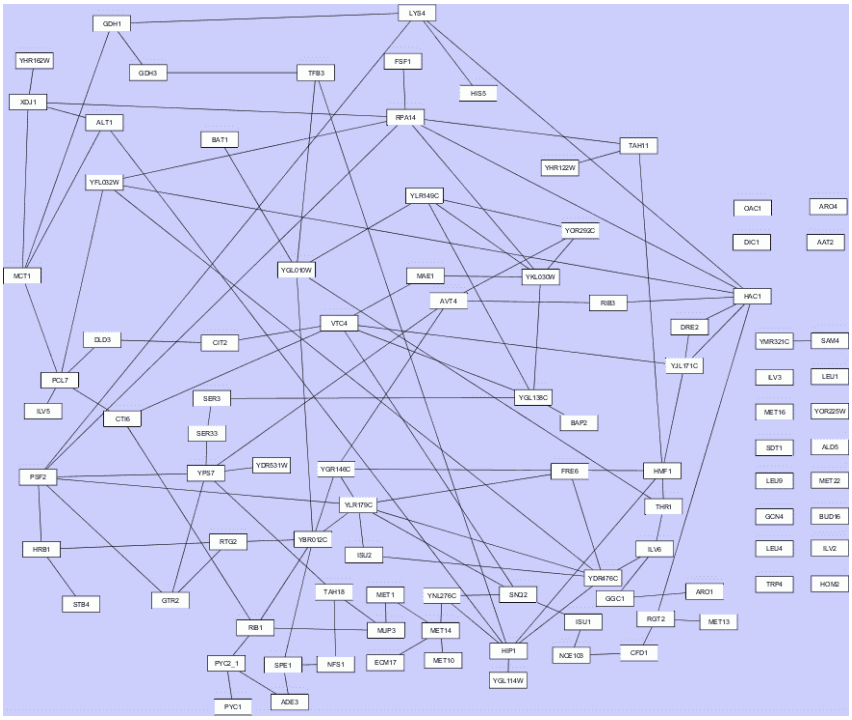
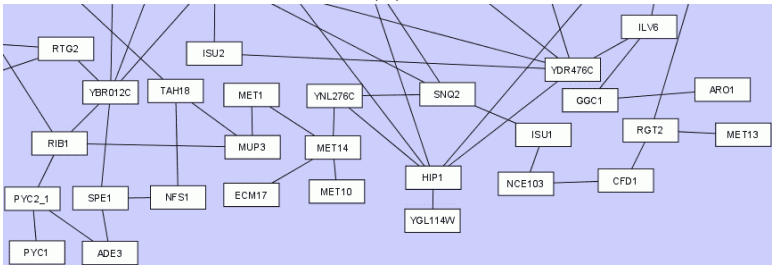



**Figure 5.7**: *Yeast metabolic pathways shown in the BioCyc Omics Viewer with highlighted edges indicating e-traits varying in the data of* Brem *et al. (2002).*

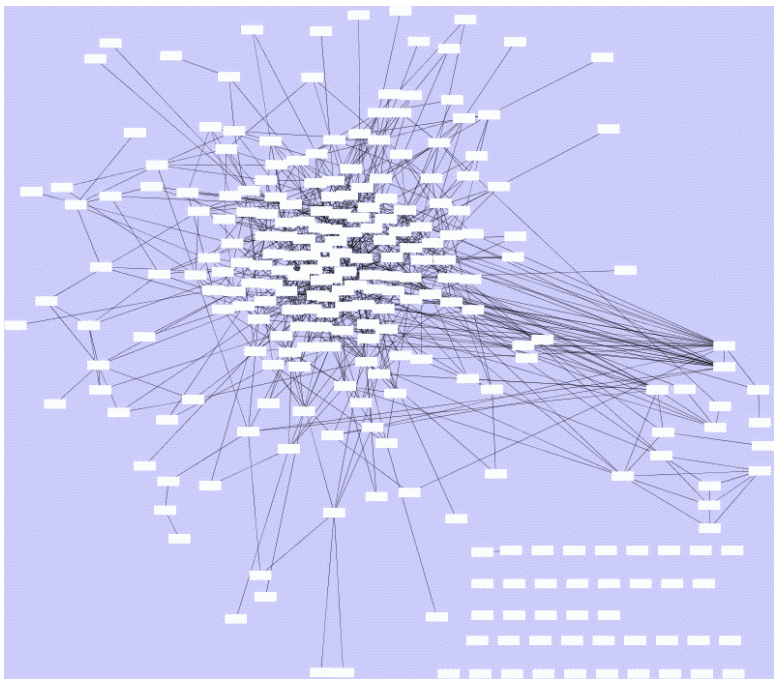



(a)

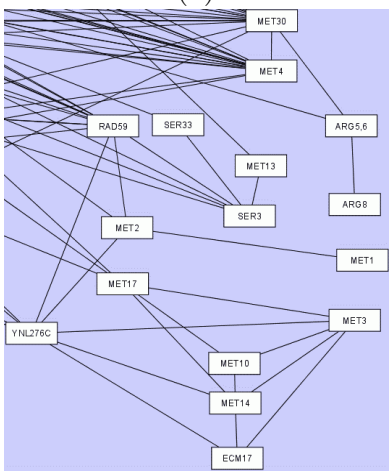



(b)

**Figure 5.8**: *Putative amino-acid- and protein-biosynthesis network constructed by GGM-eQTL from data of* Brem *et al. (2002). (a) The entire network; (b) parts that include genes in sulfate assimilation pathway.*

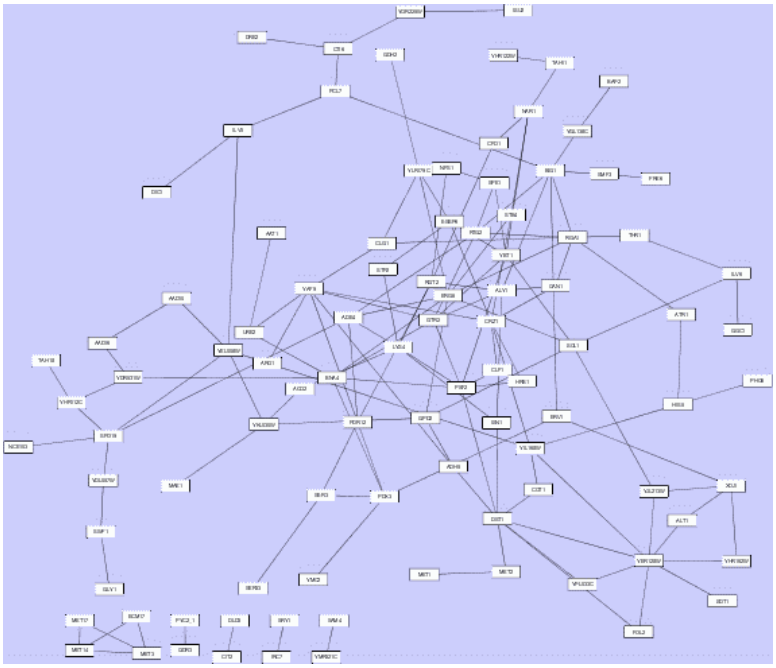


(a)

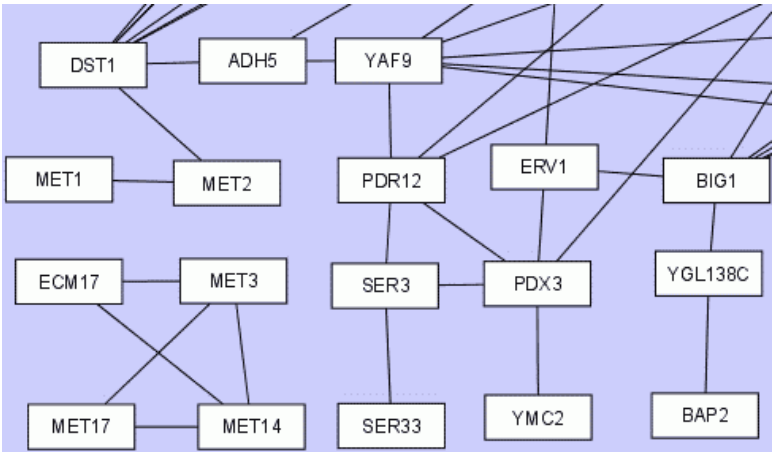



(b)

**Figure 5.9**: *Putative amino-acid-biosynthesis and -metabolism network constructed by GGM-eQTL from data of* Brem *et al. (2002). (a) The entire network; (b) parts that include genes in sulfate assimilation pathway.*



(a)




(b)

**Figure 5.10**: *Putative amino-acid-biosynthesis and -transport network constructed by GGM-eQTL from data of* Brem *et al. (2002). (a) The entire network; (b) parts that include genes in sulfate assimilation pathway.*

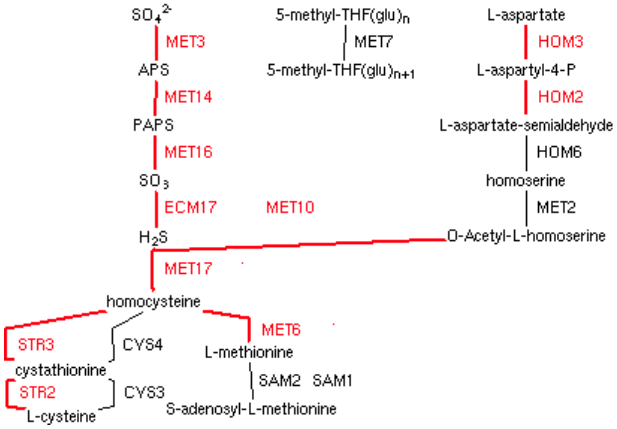



**Figure 5.11**: *Sulfate-degradation pathway retrieved from BioCyc database. Edges in red represent e-traits with* $\sigma > \log_2(0.45)$ *from data of* Brem *et al. (2002).*

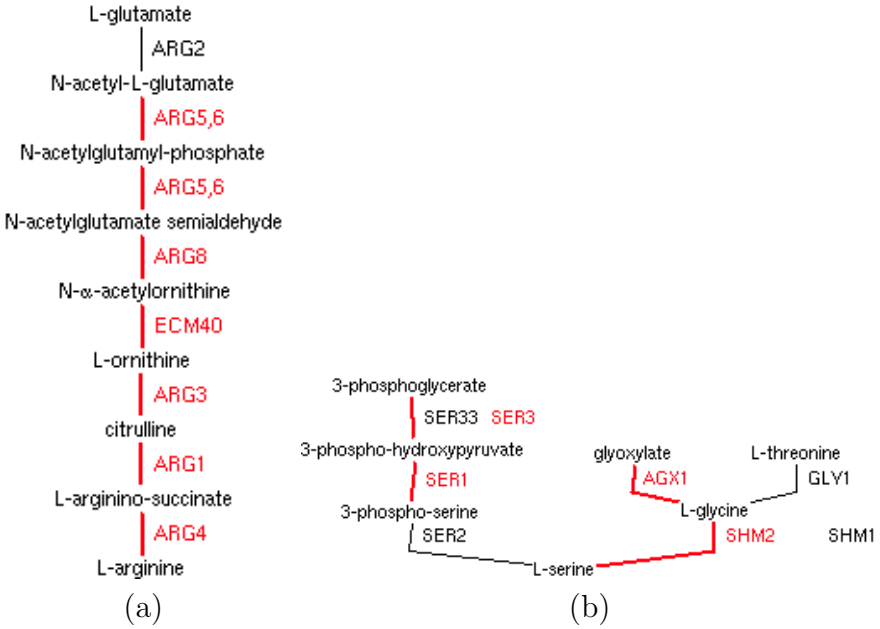



**Figure 5.12**: *Pathways of (a) arginine and (b) serine biosynthesis from 3-phosphoglycerate, retrieved from BioCyc database. Edges in red represent e-traits with* $\sigma > \log_2(0.45)$ *from data of* Brem *et al. (2002).*

some form of sulfate degradation pathway. Though genes *MET3*, *MET14*, *MET17*, and *ECM17* are linked in all three subnetworks, they are not connected as in the reference pathway shown in Figure 5.11. Genes *ARG8* and *ARG5,6*, which are involved in arginine biosynthesis (Figure 5.12 (a)), and genes *SER3* and *SER33*, which are involved in the pathway of serine biosynthesis from 3-phosphoglycerate (Figure 5.12 (b)), are linked in these subnetworks.

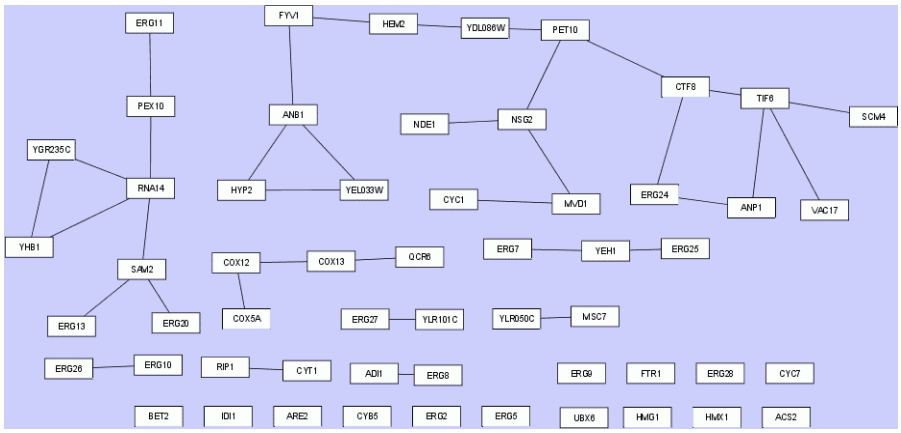



**Figure 5.13**: *Putative sterol-metabolism and electron-transport network constructed by GGM-eQTL from data of* Brem *et al. (2002).*

GGM-eQTL could reconstruct fragments of metabolic networks, despite the lack of variation in the e-trait data. For example, Figure 5.13 shows the reconstructed network (network A) from cluster 8. The gene descriptions, retrieved from SGD, suggest that the network represents a pathway for ergosterol metabolism and electron transport. The cluster contains 56 genes, connected with 36 edges. The estimate for $\lambda$ is 0.203, which is less than the global $\lambda$ of 0.406. Figure 5.14 shows the ergosterol biosynthesis pathway retrieved from BioCyc. Although only *HMG1* and *ERG3* have sufficient variation, GGM-eQTL produced fragments of the pathways. For example, *ERG11* is shown connected with *ERG13* and *ERG20*, although not in the order shown in Figure 5.14. *ERG13* is also shown connected with *MVD1*, and *ERG7* with *ERG25*. GGM-eQTL also yielded false positive edges relative to the reference figure, such as the edge between *ERG10* and *ERG26*.

As an example of constructing a network for which there is no known reference pathway,

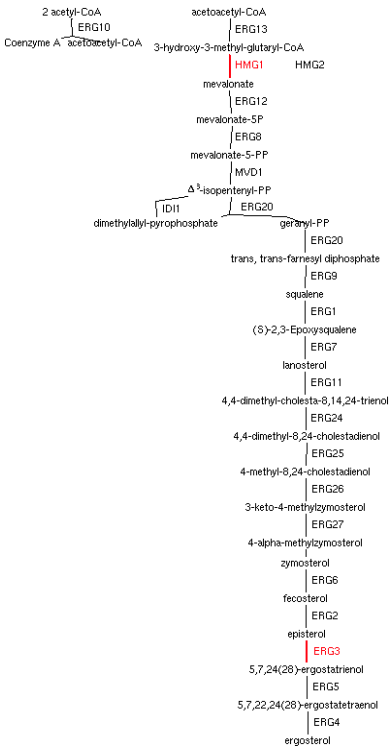



**Figure 5.14**: *Ergosterol-biosynthesis pathway retrieved from BioCyc database. Edges in red represent e-traits with* $\sigma > \log_2(0.45)$ *from data of* Brem *et al. (2002).*

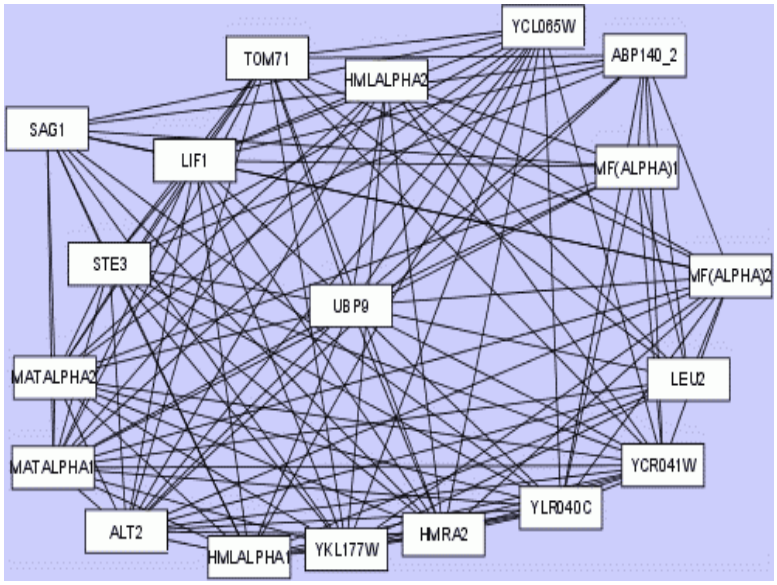



**Figure 5.15**: *Putative network for mating process constructed by GGM-eQTL from data of* Brem *et al. (2002).*

GGM-eQTL was applied to genes in cluster 1, which contains mating genes (Figure 5.15). Almost all pairs of genes are linked in the network, a relationship that seems biologically improbable. However, the low $\lambda$ value of 0.116 suggests that statistical bias is not the culprit.

In contrast to the relationships in metabolic pathways, none of the physical or genetic interactions recorded in SGD coincided with any edges in the subnetworks. None of these interactions coincided with the edges of reference pathways recorded in BioCyc database.

The bootstrap method described in section 5.5.1 for assessing edge reproducibility was performed. Owing to an error in the implementation, the results were discarded. Because the algorithm requires several days for execution, further development of this test was deferred to future work.

### 5.5.3 Discussion and conclusion

Localizing or clustering e-traits or genes provides several advantages for eQTL analysis. It saves computational time by elimination of weak gene relationships, thereby dividing the problem into manageable size. It also allows per-cluster tuning of parameter values, such as the $\lambda$ shrinkage coefficient, for network construction algorithms. This tuning may produce networks that better represent metabolic pathways because clusters may have different correlation patterns.

Conditioning on eQTL information in GGM-eQTL is also expected to yield more biologically realistic networks. The operation does this by reducing $\lambda$, meaning that less bias is introduced into the shrunk correlation matrix. Since $\lambda$ is formulated to minimize the quadratic risk function, smaller values mean that the estimated variance–covariance matrix $\mathbf{S}$ is closer to the true population variance–covariance matrix $\mathbf{\Sigma}$. The wide difference between local and global $\lambda$s suggests that the contribution of eQTL genotypic variation to e-trait variation is substantial.

The networks produced by GGM-eQTL from e-traits showing high variation across the

mapping population corresponded better to known metabolic pathways than those based on e-traits showing low variation. Both contained many spurious edges (those not recorded in reference pathways) but the latter were more fragmented. No recorded physical or genetic interactions matched edges in the subnetworks produced by GGM-eQTL. This was not surprising, since 1) interactions between proteins may not reflect genetic interactions, and 2) the reported genetic interactions were derived from experiments involving variation across environmental conditions rather than across genotypes. Such variation may induce genetic interactions that may not be expressed in the uniform environment of an eQTL experiment such as used for this study.

A rigorous comparison between the networks produced by GGM-eQTL and those produced by other network-reconstruction methods described to date is impossible. Complete networks from previous reports are unavailable or not detailed enough for extensive comparison. For example, Bing and Hoeschele (2005) showed only partial results in their Figure 3, while the complete network in their Figure 4 does not show gene names. Likewise, Keurentjes *et al.* (2007) showed only a flowering pathway in their Figure 2.

Though the bootstrap described above for edge reliability assessment seems worth investigation, it may not be useful. Bootstrap- or FDR-based cutoff values are useful only if the power of the underlying statistic is high (Zaykin and Zhivotovsky 2005). If not, such a cutoff will exclude true positives and true negatives with almost equal probability. In view of this, the ranking method described by Zaykin and Zhivotovsky (2005) may be preferable.

GGM-eQTL might be made to add direction to edges as follows. Recall matrix $\mathbf{\Omega}$, the partial-variance matrix defined in section 5.3.1, with $i^{\text{th}}$ diagonal term $\omega_{ii}$. Under the null hypothesis, the distribution of $f = \frac{\omega_{jj}}{\omega_{ii}}$ asymptotically approaches an F distribution. For an edge between e-traits $i$ and $j$, we can infer that e-trait $i$ influences e-trait $j$ if $f$ passes the $\alpha$-level critical threshold, or the reverse if $1/f$ passes the threshold. If neither condition holds, direction cannot be assigned. A variant of this approach hinted at by Schäffer and

Strimmer (2005) was tried but proved numerically unstable. Further research along this line may be fruitful.

Since network reconstruction depends on e-trait variation, future eQTL experiments should be based on crosses that maximize such variation in the pathways of interest. Perturbation in the form of appropriate stress conditions could be applied. Alternatively, a population involving multiple crosses and multiple alleles could be used.

# Chapter 6

# QGene 4, a QTL-analysis platform in Java


## Abstract

QGene 4 is an extensible high-performance QTL- and eQTL-analysis platform written in the Java programming language. The software provides many QTL mapping methods, allowing side-by-side comparisons among them. It accommodates a wide range of mating designs, can be used to simulate maps, populations, and traits and examine and transform traits, and is internationalizable and scriptable.


## 6.1 Motivation

Despite the frequent appearance of new QTL mapping methods and many claims of improvements, no recent method has been shown conclusively to be superior to other methods in precision and accuracy of QTL detection and in ease of interpretation of results. Although method papers include analyses of both real and simulated data, their authors often address only a few variations, such as of mating design, trait and marker distribution, allelic phase, or number of traits accommodated. Analysts are likely to choose a familiar method such as composite interval mapping (CIM) (Zeng 1994) in preference to newer and more powerful methods such as multiple interval mapping (MIM) (Kao *et al.* 1999). Side-by-side comparisons among QTL-mapping methods would allow analysts to examine their data

in detail with each method, thus basing their QTL findings on several different statistical approaches.

Conventional QTL software generally supports limited or no comparison among methods. Each package may rely on a different set of parameter values and analysis conditions that cannot be changed by analysts, so that a valid comparison study is difficult. Even when the source code is available, software modifications allowing such a study are typically nontrivial for programming analysts and impossible for nonprogrammers.

None of these QTL programs are built with forward extensibility, a property that allows newer modules to expand the capabilities of older ones. Such extensibility is promoted by the availability of source code or at least a well-defined application-programming interface (API). With such an API, analysts familiar with programming can build their own QTL-mapping methods without requiring much help from the original developers, who cannot be expected to keep up with the wide spectrum of new QTL methods.

Forward extensibility of software is achievable by means of a plug-in architecture. This is one in which modules, or plug-ins extending the features of the host application or of other modules, may be created by third parties and are loaded as needed. The host application alone is ready for use without programming. Such an architecture has been implemented in major Web browsers and other applications, including bioinformatics software, such as MEGA (Kumar *et al.* 2008) and Cytoscape (Shannon *et al.* 2003). Although plug-in development does require some programming, it is limited to operations close to the analyst's purposes. The host application provides and manages all other machinery required for calculation and display.

Most existing QTL software is limited to a single computer operating system (OS). Even where the source code is available, it may not be portable to other OSs. Ideally, software should be built in an OS-agnostic language, such as R or Java.

QGene 4 (Joehanes and Nelson 2008), a complete rewrite of an earlier application

(Nelson 1997), is designed to address these issues. Its features are described in the following sections.

## 6.2 Features for analysts

The methods described in Chapters 2–4 plus inclusive composite interval mapping (ICIM) (Li *et al.* 2007) have been implemented in QGene as of version 4.3. These methods can be invoked from the QTL analysis window, as shown in Figure 6.1. In QGene, display updates, such as after selection of chromosomes, traits, analyses, cofactors for QTL methods such as CIM, and resizing, are performed dynamically, allowing the analysts to compare QTL profiles of different methods quickly. Permutation analysis is implemented in QGene for all QTL methods. QGene also allows the creation of publication-ready figures in a variety of formats.

QGene accommodates a wide range of mating designs, encoded as strings, each character in which represents a breeding operation starting from the $F_1$ generation. Any sequence of *b*, *d*, *s*, and *i* or their upper-case counterparts may be provided as a mating string. These letters refer to *b*ackcrossing, *d*ouble-haploid creation, *s*elfing, and random *i*ntercrossing. A $BC_1F_1$ design, for example, is specified by *b*; an $F_2$ by *s*; an $F_3$ by *ss*; and a series of three backcrosses followed by a selfing by *bbbs*. The lone letter *r* refers to a standard recombinant-inbred progeny, created by multiple-generation selfing but possibly retaining some heterozygosity.

Trait analysis in QGene allows analysts to inspect trait normality and trait correlation matrices, as shown in Figure 6.2. QGene also allows transformation of traits, simulation of missing trait data, and regression of one trait on others.

QGene can be used to simulate maps, populations with mating designs as mentioned above, and traits, including multiple correlated traits. Most parameters are adjustable, such as percentage of missing genotype data, QTL positions and trait-specific effects, and the means and distributions of traits.

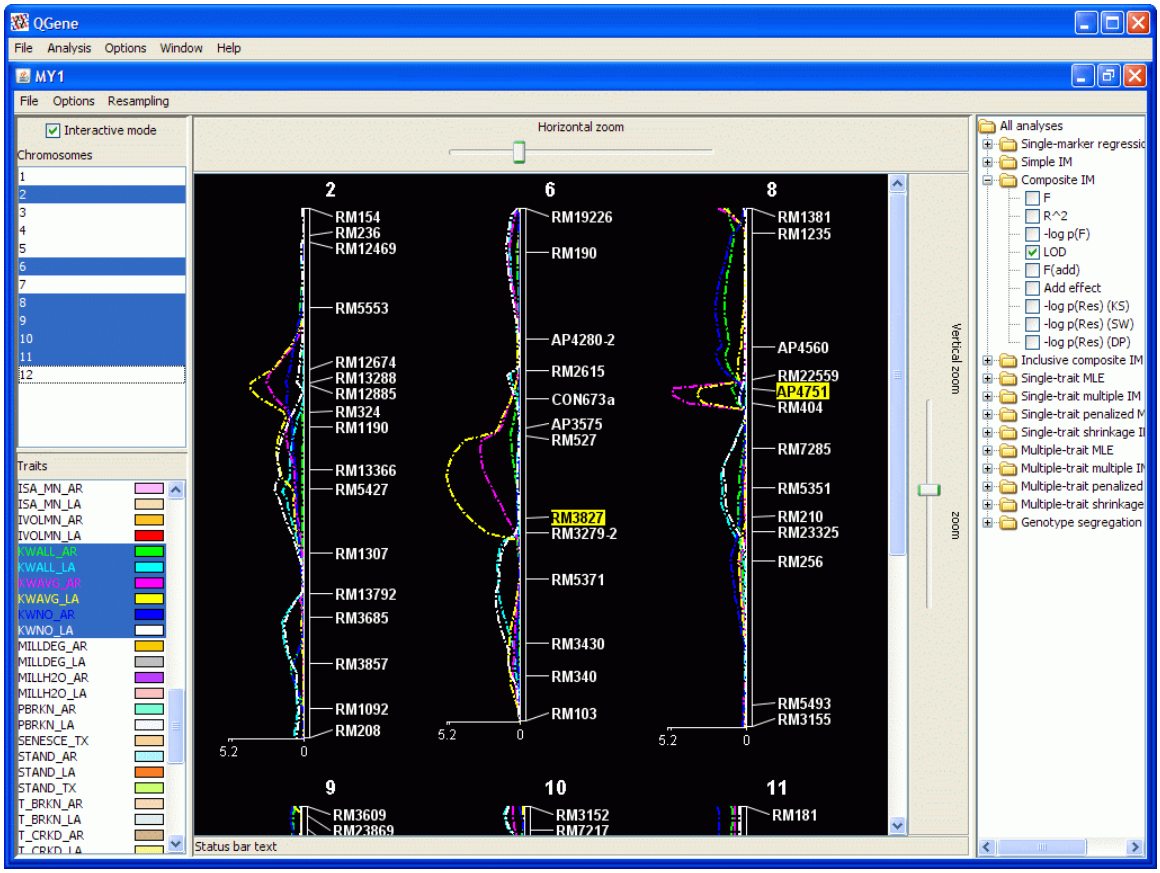



**Figure 6.1**: *QTL analysis window in QGene accommodates comparisons among multiple QTL-analysis methods.*

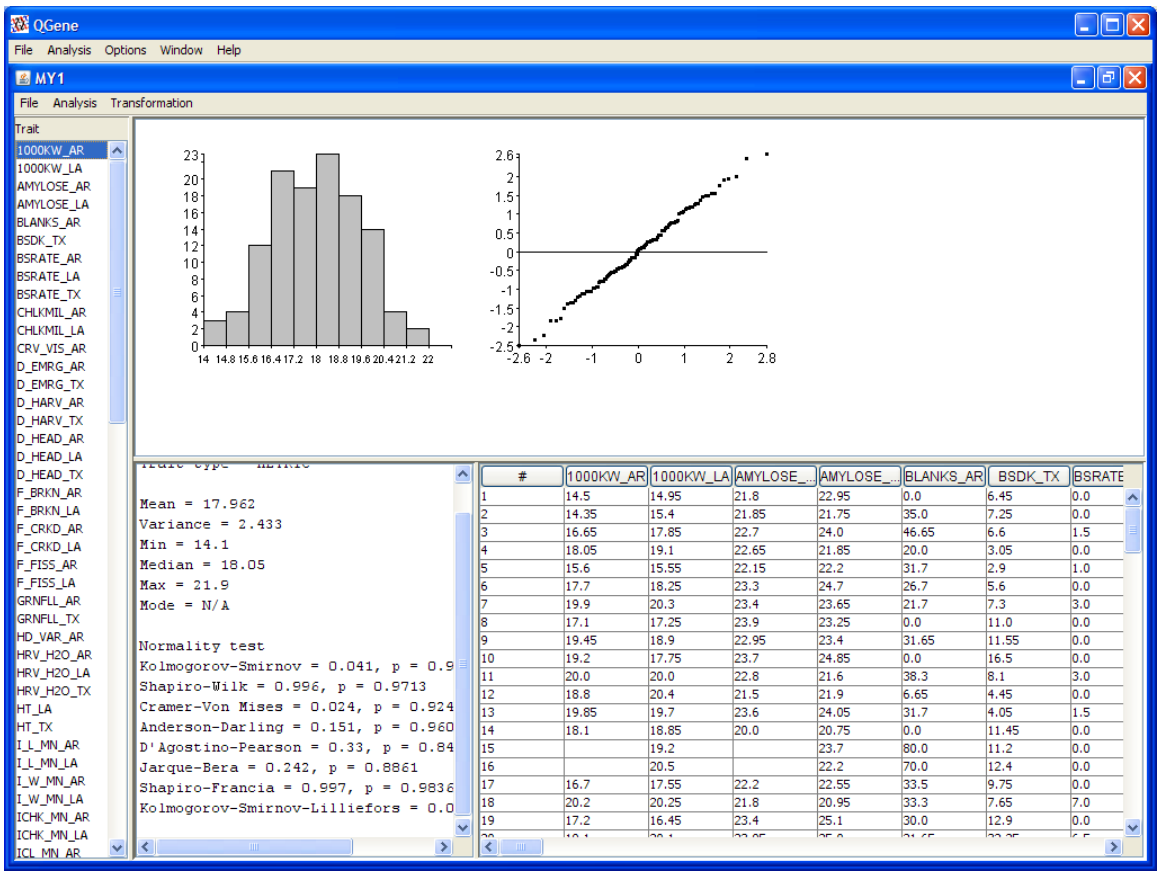



**Figure 6.2**: *Trait analysis window in QGene.*

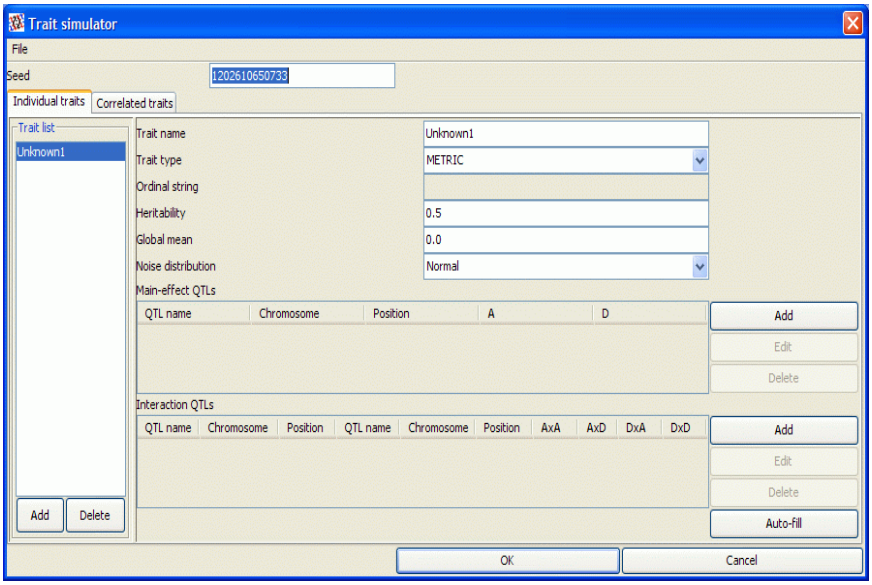



**Figure 6.3**: *Trait simulation window in QGene.*

## 6.3 Architecture

QGene 4 is implemented in the popular, multi-platform, high-performance Java language. It is powered by a multi-threaded computation engine to speed up computations.

QGene 4 is designed with a plug-in architecture, with almost all components written as plug-ins. For example, if an analysis method is implemented correctly as a plug-in, QGene's permutation analysis plug-in, which computes the threshold above which a QTL can be declared significant (Churchill and Doerge 1994), can be used for that method without extra effort. If the method employs the genotype probability distributions of QTLs and missing markers (Jiang and Zeng 1997), it can accommodate missing markers and all mating designs that QGene accepts. Minimal knowledge of QGene internal machinery is required, though the software is also open-source. While the easiest plug-ins to develop will be for QTL mapping methods, graphical or GUI plug-ins are also possible for more advanced developers.

Almost all QGene components are coded as plug-ins. At load time, a plug-in manager scans for loadable plug-ins in `.jar` format in the `lib` subdirectory of QGene installation directory and the `.qgene` subdirectory of the user's home directory. Each plug-in can interact with other plug-ins according to the QGene API specification. The plug-in architecture in QGene is built upon the class loader feature in Java, which allows dynamic loading and invocation of precompiled Java classes.

There are three main plug-in categories in QGene: general analysis, statistical calculation, and rendering. Statistical-calculation plug-ins are used for QTL mapping computation. Rendering plug-ins, such as contour and heat-map plug-ins, are used for rendering computation results to the screen. General analysis plug-ins, such as the main QTL window, eQTL, and trait analysis plug-ins, are used for providing new types of analysis.

It is possible to create new plug-in categories in QGene. Plug-in categories are built by extension of (creation of classes that inherit from) the `QPlugin` and `QPluginGroup` classes,

as shown in Listing 6.1. Here, the plug-in base class is `MyNewPlugin` and the plug-in category is `MyNewPluginGroup`.

**Listing 6.1**: *Creating a new plug-in category*

```
1  import qplugin.*;
2  public abstract class MyNewPlugin extends QPlugin {
3    public final Class<QStatsPluginGroup> getPluginGroupClass()
4    { return MyNewPluginGroup.class; }
5    // Other abstract methods may be added
6  }
7
8  public class MyNewPluginGroup extends QPluginGroup {
9    public Class<MyNewPlugin> getPluginBaseClass()
10   { return MyNewPlugin.class; }
11
12   public String getPluginGroupName()
13   { return "MyPluginGroup"; }
14 }
```

A new class must be created by extension of the plug-in base class and implementation of all abstract methods in order to create a new plugin. For example, a new QTL-mapping plug-in must inherit from the `QGeneticStatsPlugin` class. Abstract methods `calculate` and `getOutputType` must be implemented. For general QTL-mapping plug-ins, the `getOutputType` method must return the class object of `QMultiChromValues[]`. An example is shown in Listing 6.2.

**Listing 6.2**: *Creating a new QTL mapping plug-in*

```
1  import jama.Matrix;
2  import java.util.List;
3  import qgenetics.*;
4  import qplugin.stats.*;
5  import qplugin.stats.expectation.*;
6  import static qstats.QStatsUtils.kUndefinedValue;
7
8  public class QExamplePlugin extends QGeneticStatsPlugin<QMultiChromValues[],
       QMultiChromInteractionValues[]>
9  {
10   protected IStatistic<? extends QStatsPlugin<QMultiChromValues[]>>[] mStats =
         createOutputStats();
11
12   /**
13    * Return true if the input data is compatible with this plug-in analysis.
14    * The input data is mData. Although mData is declared as type IPluginData,
15    * this has already been verified by QGeneticStatsPlugin
16    */
17   @Override
18   public boolean initData()
19   {
20     boolean retVal = super.initData();
21     if (!retVal)
22       return false;
23     if (mExpectation == null)
24       mExpectation = new QMarkovChainExpectationPlugin();
```

```
25      mExpectation.purgeCache();
26      mExpectation.initData(mData);
27      // For example, if this analysis tackles only F2, return false if the mating string is
            not of type F2.
28      // if (!mPopn.getMatingParams().getMatingString().equalsIgnoreCase("S")) // $NON-NLS-1$
29      //  return false;
30      // Or, alternatively:
31      // if (!QGenetics.Mating.QMatingUtil.isF2(mPopn.getMatingParams()) return false;
32      return true;
33    }
34
35    /**
36     * Do your computation here.
37     * @param chroms indicates the list of chromosomes selected in the GUI
38     * @param traits indicates the list of traits selected in the GUI
39     * @param stat indicates the statistic chosen in the GUI.
40     * @return the result for the specific chosen chromosomes, traits, and statistic.
41     */
42    @Override
43    public QMultiChromValues[] calculate(List<QChromosome> chroms, List<QTrait> traits,
          IStatistic<? extends QStatsPlugin<QMultiChromValues[]>> stat)
44    {
45      // mMap is the map object
46      // mPopn is the population object
47      // Both of these are already set for you.
48
49      // Here's an example of iterating over the loci available in the map
50      for (QChromosome chrom: mMap.getChromosomes())
51      {
52        int chromIndex = chrom.getIndex(); // the global index of the chromosome
53        assert (chromIndex >= 0);
54        // chromIndex will have the true index.
55        for (QLocus locus: chrom.getAllLoci())
56        {
57          int
58            locusAbsolutePosInMM = locus.getPosition(), // in milliMorgans (mM) from the
                  start of the chromosome
59            chromLengthInMM = chrom.getLength();
60          assert (locusAbsolutePosInMM >= 0 && locusAbsolutePosInMM <= chromLengthInMM);
61          // Here's an example of how to get the genotype array
62          // The order is sorted according to the individuals, which means that
63          // genotypeData[i] will always refer to the genotype data of i-th individual of
                locus <tt>locus</tt>
64          QGType[] genotypeData = mPopn.extractGenotypeData(locus);
65          assert (genotypeData != null);
66        }
67        QGType[][] genotypeData = mPopn.extractGenotypeData(chrom); // this is for one
              chromosome
68        assert (genotypeData != null);
69      }
70      // WARNING: the order of the chromosomes in chroms is not necessarily the same as the
            order in mMap.getChromosomes()
71      // For example: In chroms, we may have chromosome 1, 4, and 7 chosen; while
            mMap.getChromosomes() will
72      // always return all chromosomes in the map (i.e. 1, 2, 3, ..., 7). So if your
            algorithm relies on indices of arrays,
73      // you'd better take note.
74
75      // We will need chrom.getIndex() to get around it. However, any indexing mechanisms is
            brittle against
76      // modifications. That is, if some module modifies a chromosome (either by deleting or
            adding some loci
77      // or by some other means), then the chromosome index is no longer valid. A better way
            to deal with this
```

```
78      // is to use some hashing mechanism that compares the chromosome names. This may incur
            some speed penalty, but
79      // it is far more stable and more portable than an indexing mechanism.
80
81      // Here's an example of how to get trait data
82      for (QTrait trait : mPopn.getTraits())
83      {
84        double[] traitData = trait.getTraitData();
85        assert (traitData != null); // this will always be true
86        for (double traitVal : traitData)
87        {
88          if (traitVal == kUndefinedValue)
89            continue;
90        }
91      }
92
93      // Here's an example of how to obtain genotype expectation
94      // NOTE: If you need genotype expectation, you must return true in
            needsExpectationCalculator().
95      // Otherwise, this code will throw a null-pointer exception
96      mExpectation.setXInterval(10); // say, if you don't like the 20 mM default distance
97      QExpectationResult expResult = mExpectation.calculate(chroms, traits, null);
98      assert (expResult != null);
99      double[][][] // arranged in chrom x loci x individuals
100       addExp = expResult.mLocusExpectations[0],
101       qtlAddExp = expResult.mQTLExpectations[0],
102       domExp = expResult.mLocusExpectations[1],
103       qtlDomExp = expResult.mQTLExpectations[1];
104     Matrix[][][] // arranged in chrom x loci x individuals
105       locusCPD = expResult.mLocusCPDs,
106       qtlCPD = expResult.mQTLCPDs;
107     // Note: The chrom order will be according to the chroms list, not the global
            chromosome order!
108     assert (addExp != qtlAddExp && domExp != qtlDomExp && qtlCPD != null && locusCPD !=
            null);
109
110     // Do your calculation here, but do NOT return null. This is only an example.
111     return null;
112   }
113
114   /**
115    * If you have multiple statistics that can be computed by this plug-in, return an array
           of IStatistic
116    * instead. If you return null, then you have only one statistic in this plug-in.
117    */
118   @Override
119   public IStatistic<? extends QStatsPlugin<QMultiChromValues[]>>[] getAvailableStatistics()
120   {
121     if (mStats == null)
122       mStats = createOutputStats();
123     return mStats;
124   }
125
126   protected IStatistic<? extends QStatsPlugin<QMultiChromValues[]>>[] createOutputStats()
127   {
128     return QDefaultStatistic.createStatisticObjects(this, new String[] { "Stats1",
            "Stats2", "Stats3" });
129     // Or
130     // return QDefaultStatistic.createStatisticObjects(this, new String[] { "F", "R^2",
            "LOD" });
131   }
132
133   /**
134    * Whether or not the plug-in needs an expectation engine
135    * @return true if you need genotype expectation calculator
```

```
136     */
137    @Override
138    public boolean needsExpectationCalculator()
139    { return true; }
140
141    /**
142     * Modify this method only if you want to make a statistic plug-in that is being used by
143     * some analyses other than QTL mapping analysis. This is because the QTL mapping window
           will
144     * use this to automatically discover statistics plug-ins compatible with its input type.
145     */
146    @Override
147    public Class<QMultiChromValues[]> getOutputType()
148    { return QMultiChromValues[].class; }
149
150    /**
151     * This method is just like getOutputType() but it's used for interaction analysis.
152     * @see qplugin.stats.QGeneticStatsPlugin#getInteractionOutputType()
153     */
154    @Override
155    public Class<QMultiChromInteractionValues[]> getInteractionOutputType()
156    { return QMultiChromInteractionValues[].class; }
157  }
```

## 6.4 Other features

QGene is distributed as six Java .jar libraries, of which *qgene.jar* and *qgenerics.jar* were written by the author. The program has been written to support internationalization of its interface via translation of a set of text resource files into any desired language, though to date only an English-language version is available.

QGene can be run from another Java program, a convenient feature that allows execution of lengthy simulation experiments. The short example shown here (with import statements removed) loads an input file, runs single-trait MIM, and saves the text and graphical output to files.

**Listing 6.3**: *A QGene script running single-trait MIM*

```
1  public class MIMmer {
2      public static void main(String[] args) {
3        QPluginFactory.registerDefaultPlugins();
4        try {
5          QGeneticData data = new QQDFPlugin().load(new FileRead-er("data.qdf"));
6          List<QChromosome> chroms = data.mMap.getChromosomes();
7          List<QTrait> allTraits = data.mPopn.getTraits();
8          List<QTrait> traits = new ArrayList<QTrait>();
9          traits.add(allTraits.get(0)); // Choose only the first trait
10         QSingleTraitMIMPlugin mim = new QSingleTraitMIMPlugin();
11         mim.initData(data);
12         QMultiChromValues[] results = mim.calculate
13           (chroms, traits, mim.getAvailableStatistics()[0]);
```

```
14          StringBuilder textOutput = new StringBuilder
15            (QMultiChromValues.createReportTable(results).toString());
16          textOutput.append(mim.getAdditionalCalculationInfo());
17          QFileUtils.writeText(textOutput.toString(),"output.txt");
18          QDualSerialContourRendererPlugin renderer =
19            new QDualSerialContourRendererPlugin();
20          renderer.setData(results);
21          renderer.saveImage(EImageTypes.PNG, "contour.png");
22          } catch (Exception e) {
23          e.printStackTrace();
24        }
25      }
26  }
```

## 6.5 Future development

With QGene modular architecture, a capable analyst can build custom plug-ins for additional analyses currently not available in QGene, such as analyses of epistatic interactions, fixed- and random-effect covariates, and outcross populations. The author would like to add some of these analyses in the future, as time permits.

# Bibliography


Agresti, A., 2002 *Categorical Data Analysis,* 2nd ed. John Wiley and Sons, Hoboken, New Jersey.

Baldwin, N. E., E. J. Chesler, S. Kirov, M. A. Langston, J. R. Snoddy, R. W. Williams, and B. Zhang, 2005 Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks. Journal of Biomedical Biotechnology **2005**: 172–180.

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

Bezdek, J., 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Bing, N. and I. Hoeschele, 2005 Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics **170**: 533–542.

Boer, M. P., C. J. F. ter Braak, and R. Jansen, 2002 A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **162**: 951–960.

Box, G. E. P., 1953 Non-normality and tests on variances. Biometrika **40**: 318–335.

Breitling, R., A. Amtmann, and P. Herzyk, 2004 Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics **5**: 1–8.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, and R. C. Jansen, 2008 Genetical genomics: spotlight on QTL hotspots. PLOS Genetics **4**: 1–4.

Brem, R., G. Yvert, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296**: 752–755.

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

Chickering, D. M., D. Geiger, and D. Heckerman, 1994 Learning Bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft.

Chickering, D. M., D. Heckerman, and C. Meek, 2004 Large-sample learning Bayesian networks is NP-hard. Journal of Machine Learning Research **5**: 1287–1330.

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

Coffman, C. J., R. W. Doerge, K. L. Simonsen, K. M. Nichols, C. K. Duarte, R. D. Wolfinger, and L. M. McIntyre, 2005 Model selection in binary trait locus mapping. Genetics **170**: 1281–1297.

Cooper, G. F. and E. Herskovitz, 1992 A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9**: 309–347.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 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.

Dojer, N., 2006 Learning Bayesian networks does not have to be NP-hard. Lecture Notes in Computer Science **4162**: 305–314.

Dwight, S. S., M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry, 2002 Saccharomyces genome database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Research **30**: 69–72.

Eisen, M. B., P. Spellman, P. Brown, and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences USA **95**: 14863–14868.

Fahmeir, L. and G. Tutz, 2001 *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, second edition.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er, 2000 Using Bayesian networks to analyze expression data. Journal of Computational Biology **7**: 601–620.

Gasch, A. and M. B. Eisen, 2002 Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Biology **3(11)**, http://rana.lbl.gov/FuzzyK/Home.htm.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003 *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, Florida, second edition.

Gene Ontology Consortium, 2000 Gene ontology: tool for the unification of biology. Nature Genetics **25**: 25–29.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrics **82**: 711–732.

Guo, Z., R. Joehanes, and J. C. Nelson, 2007 Shrinkage interval mapping for QTL and QTL epistasis analysis in line crosses. Unpublished manuscript.

Hackett, C. A. and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51**: 1252–1263.

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.

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.

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

Heckerman, D., 1998 A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, edited by M. I. Jordan, chapter 7, pp. 301–354, The MIT Press.

Hoerl, A. E., 1962 Application of ridge analysis to regression problems. Chemical Engineering Progress **58**: 54–59.

Hoerl, R. W., J. H. Schuenemeyer, and A. E. Hoerl, 1986 A simulation of biased estimation and subset selection regression techniques. Technometrics **28**: 369–380.

Jansen, R. and J.-P. Nap, 2001 Genetical genomics: the added value from segregation. Trends in Genetics **17**: 388–391.

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 P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136**: 1447–1455.

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

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.

Jiang, D., C. Tang, and A. Zhang, 2004 Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering **16**: 1370–1386.

Joehanes, R., 2009 *Multiple-trait multiple-interval mapping of quantitative-trait loci*. Master's thesis, Kansas State University, Manhattan, Kansas, http://hdl.handle.net/2097/1605.

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

Jørgensen, B., 1983 Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika **70**: 19–28.

Kao, C.-H. and Z.-B. Zeng, 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. and Z.-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160**: 1243–1261.

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

Karp, P. D., C. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas, 2005 Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research **19**: 6083–89.

Karp, P. D., S. Paley, and P. Romero, 2002 The pathway tools software. Bioinformatics **18**: S1–S8.

Keurentjes, J. J. B., J. Fu, C. H. R. de Vos, A. Lommen, R. D. Hall, R. J. Bino, L. H. W. van der Plas, R. Jansen, D. Vreugdenhil, and M. Koorneef, 2006 The genetics of plant metabolism. Nature Genetics **38**: 842–849.

Keurentjes, J. J. B., J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. Jansen, 2007 Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences USA **104**: 1708–1713.

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

Kulp, D. C. and M. Jagalur, 2006 Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics **7**: 1–12.

Kumar, S., M. Nei, J. Dudley, and K. Tamura, 2008 MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics **9**: 299–306.

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

Li, H., G. Ye, and J. Wang, 2007 A modified algorithm for the improvement of composite interval mapping. Genetics **175**: 361–374.

Li, J., S. Wang, and Z.-B. Zeng, 2006 Multiple-interval mapping for ordinal traits. Genetics **173**: 1649–1663.

Li, R., S.-W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen, and G. A. Churchill, 2006 Structural model analysis of multiple quantitative traits. PLOS Genetics **2**: e114.

Liu, B., A. de la Fuente, and I. Hoeschele, 2008 Gene network inference via structural equation modeling in genetical genomics experiments. Genetics **178**: 1763–1776.

Lum, P. Y., Y. Chen, J. Zhu, J. Lamb, S. Melmed, S. Wang, T. A. Drake, A. J. Lusis, and E. E. Schadt, 2006 Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. Journal of Neurochemistry **s1**: 50–62.

Mähler, M., C. Most, S. Schmidtke, J. P. Sundberg, R. Li, H. J. Hedrich, and G. A. Churchill, 2002 Genetics of colitis susceptibility in IL-10-deficient mice: backcross versus F2 results contrasted by principal component analysis. Genomics **80**: 274–282.

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

Meng, X.-L. and D. B. Rubin, 1993 Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrics **80**: 267–278.

Murphy, K. and S. Mian, 1999 Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA.

Neapolitan, R. E., 2003 *Learning Bayesian networks (Artificial Intelligence)*. Prentice Hall, New York.

Nelder, J. and R. Wedderburn, 1972 Generalized linear models. Journal of the Royal Statistical Society B **135**: 370–384.

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

Opgen-Rhein, R. and K. Strimmer, 2007 Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Statistical Applications in Genetics and Molecular Biology **6**.

Pearl, J., 1988 *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kauffman, New York.

Raftery, A. E. and S. M. Lewis, 1995 The number of iterations, convergence diagnostics, and general Metropolis algorithms. In *In Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter, et. al*, pp. 115–130, Chapman and Hall.

Rao, S. and X. Li, 2000 Strategies for mapping of categorical traits. Genetica **109**: 183–197.

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

Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis, 2005 An integrative genomics approach to infer causal associations between expression and disease. Nature Genetics **37**: 710–717.

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

Schäffer, J. and K. Strimmer, 2005 A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology **4**: 1–30.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genetical Research **152**: 1203–1216.

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

Søballe, B. and R. K. Poole, 1999 Microbial ubiquinones: multiple roles in respiration, gene regulation and oxidative stress management. Microbiology **145**: 1817–1830.

Spiegelhalter, D., A. Dawid, S. Lauritzen, and R. Cowell, 1993 Bayesian analysis in expert systems. Statistical Science **8**: 219–282.

Steinmetz, L. M., H. Sinha, D. R. Richards, J. I. Spiegelman, P. J. Oefner, J. H. Cusker, and R. W. Davis, 2002 Dissecting the architecture of a quantitative trait locus in yeast. Nature **416**: 326–330.

Stephens, D. A. and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54**: 1334–1347.

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

Wang, H., Y. M. Zhang, X. Li, G. L. Masinde, S. Mohan, D. J. Baylink, and S. Xu, 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics **170**: 465–480.

Wang, S., C. J. Basten, and Z.-B. Zeng, 2007 Windows QTL Cartographer 2.5 http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Watts, J. A., M. Morley, J. T. Burdick, J. L. Fiori, W. J. Ewens, R. S. Spielman, and V. G. Cheung, 2002 Gene expression phenotype in heterozygous carriers of *ataxia telangiectasia*. American Journal of Human Genetics **71**: 791–800.

West, M. A. L., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore, R. W. Doerge, and D. A. St.Clair, 2007 Global eQTL mapping reveals the complex genetic architecture of transcript level variation in *Arabidopsis*. Genetics **175**: 1441–1450.

Wittenburg, D., V. Guiard, F. Liese, and N. Reinsch, 2007 Linear and generalized linear models for the detection of QTL effects on within-subject variability. Genetical Research, Cambridge **89**: 245–257.

Woodbury, M. A., 1950 Inverting modified matrices. Technical Report MR0038136, Statistical Research Group, Princeton University, Princeton, NJ.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics **163**: 789–801.

Xu, S., N. Yi, D. Burke, A. Galecki, and R. A. Miller, 2003 An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. Genetical Research, Cambridge **82**: 127–138.

Yi, N., S. Banerjee, D. Pomp, and B. S. Yandell, 2007 Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. Genetics **176**: 1855–1864.

Yi, N. and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155**: 1391–1403.

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

Yvert, G., R. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak, 2003 *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics **35**: 57–64.

Zaykin, D. V. and L. A. Zhivotovsky, 2005 Ranks of genuine associations in whole-genome scans. Genetics **171**: 813–823.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136**: 1457–1468.

Zhang, Y.-M., 2006 Shrinkage estimation method for mapping multiple quantitative trait loci. Acta Genetica Sinica **33**: 861–869.

Zhang, Y.-M. and S. Xu, 2005 A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity **95**: 96–104.

Zhu, C., J. Huang, and Y.-M. Zhang, 2007 Mapping binary trait loci in the $F_{2:3}$ design. Heredity **98**: 337–344.

Zhu, J., P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt, 2004 An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenetic and Genome Research **105**: 363–374.

# Appendix A

# Genes found by Fuzzy-K clustering

Gene descriptions were retrieved from the Saccharomyces Genome Database (SGD) (Dwight *et al.* 2002) on August 24, 2009. Uppercase letters in a gene's short description, if one appears, represents the abbreviation used for the gene name.

Table A.1: **Genes in a putative sterol-metabolism and electron-transport cluster formed by JFuzzy-K from data of Brem *et al.* (2002).**

| Name | Short description | Full description |
|---|---|---|
| ACS2 | Acetyl CoA Synthetase | Acetyl-coA synthetase isoform which, along with Acs1p, is the nuclear source of acetyl-coA for histone acetylation; mutants affect global transcription; required for growth on glucose; expressed under anaerobic conditions |
| ANP1 | ANP and osmotic sensitive | Subunit of the alpha-1,6 mannosyltransferase complex; type II membrane protein; has a role in retention of glycosyltransferases in the Golgi; involved in osmotic sensitivity and resistance to aminonitrophenyl propanediol |
| ARE2 | Acyl-coenzyme A: cholesterol acyl transferase-Related Enzyme | Acyl-CoA:sterol acyltransferase, isozyme of Are1p; endoplasmic reticulum enzyme that contributes the major sterol esterification activity in the presence of oxygen |
| ATG33 | AuTophaGy related | Putative protein of unknown function with similarity to SCM4; green fluorescent protein (GFP)-fusion protein localizes to mitochondria; YLR356W is not an essential gene |
| BET2 | Blocked Early in Transport | Beta subunit of Type II geranylgeranyltransferase required for vesicular transport between the endoplasmic reticulum and the Golgi; provides a membrane attachment moiety to Rab-like proteins Ypt1p and Sec4p |

*continued on the next page*

**Genes in a putative sterol-metabolism and electron-transport cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| COX5A | Cytochrome c OXidase | Subunit Va of cytochrome c oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain; predominantly expressed during aerobic growth while its isoform Vb (Cox5Bp) is expressed during anaerobic growth |
| COX12 | Cytochrome c OXidase | Subunit VIb of cytochrome c oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain; required for assembly of cytochrome c oxidase but not required for activity after assembly; phosphorylated |
| COX13 | Cytochrome c OXidase | Subunit VIa of cytochrome c oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain; not essential for cytochrome c oxidase activity but may modulate activity in response to ATP |
| CTF8 | Chromosome Transmission Fidelity | Subunit of a complex with Ctf18p that shares some subunits with Replication Factor C and is required for sister chromatid cohesion |
| CYB5 | CYtochrome B | Cytochrome b5, involved in the sterol and lipid biosynthesis pathways; acts as an electron donor to support sterol C5-6 desaturation |
| CYC1 | Cytochrome c | Cytochrome c, isoform 1, electron carrier of the mitochondrial intermembrane space that transfers electrons from ubiquinone-cytochrome c oxidoreductase to cytochrome c oxidase during cellular respiration |
| CYC7 | Cytochrome c | Cytochrome c isoform 2, expressed under hypoxic conditions; electron carrier of the mitochondrial intermembrane space that transfers electrons from ubiquinone-cytochrome c oxidoreductase to cytochrome c oxidase during cellular respiration |
| CYT1 | Cytochrome c1 | Cytochrome c1, component of the mitochondrial respiratory chain; expression is regulated by the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex |
| ERG2 | ERGosterol biosynthesis | C-8 sterol isomerase, catalyzes the isomerization of the delta-8 double bond to the delta-7 position at an intermediate step in ergosterol biosynthesis |
| ERG5 | ERGosterol biosynthesis | C-22 sterol desaturase, a cytochrome P450 enzyme that catalyzes the formation of the C-22(23) double bond in the sterol side chain in ergosterol biosynthesis; may be a target of azole antifungal drugs |
| ERG7 | ERGosterol biosynthesis | Lanosterol synthase, an essential enzyme that catalyzes the cyclization of squalene 2,3-epoxide, a step in ergosterol biosynthesis |
| ERG8 | ERGosterol biosynthesis | Phosphomevalonate kinase, an essential cytosolic enzyme that acts in the biosynthesis of isoprenoids and sterols, including ergosterol, from mevalonate |
| ERG9 | ERGosterol biosynthesis | Farnesyl-diphosphate farnesyl transferase (squalene synthase), joins two farnesyl pyrophosphate moieties to form squalene in the sterol biosynthesis pathway |
| ERG10 | ERGosterol biosynthesis | Acetyl-CoA C-acetyltransferase (acetoacetyl-CoA thiolase), cytosolic enzyme that transfers an acetyl group from one acetyl-CoA molecule to another, forming acetoacetyl-CoA; involved in the first step in mevalonate biosynthesis |

*continued on the next page*

**Genes in a putative sterol-metabolism and electron-transport cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| ERG11 | ERGosterol biosynthesis | Lanosterol 14-alpha-demethylase, catalyzes the C-14 demethylation of lanosterol to form 4,4"-dimethyl cholesta-8,14,24-triene-3-beta-ol in the ergosterol biosynthesis pathway; member of the cytochrome P450 family |
| ERG20 | ERGosterol biosynthesis | Farnesyl pyrophosphate synthetase, has both dimethylallyltranstransferase and geranyltranstransferase activities; catalyzes the formation of C15 farnesyl pyrophosphate units for isoprenoid and sterol biosynthesis |
| ERG24 | ERGosterol biosynthesis | C-14 sterol reductase, acts in ergosterol biosynthesis; mutants accumulate the abnormal sterol ignosterol (ergosta-8,14 dienol), and are viable under anaerobic growth conditions but inviable on rich medium under aerobic conditions |
| ERG25 | ERGosterol biosynthesis | C-4 methyl sterol oxidase, catalyzes the first of three steps required to remove two C-4 methyl groups from an intermediate in ergosterol biosynthesis; mutants accumulate the sterol intermediate 4,4-dimethylzymosterol |
| FTH1 | FTS3 Homolog | Putative high affinity iron transporter involved in transport of intravacuolar stores of iron; forms complex with Fet5p; expression is regulated by iron; proposed to play indirect role in endocytosis |
| FTR1 | Fe TRansporter | High affinity iron permease involved in the transport of iron across the plasma membrane; forms complex with Fet3p; expression is regulated by iron |
| HAP2 | Heme Activator Protein | Subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; contains sequences sufficient for both complex assembly and DNA binding |
| HEM2 | HEMe biosynthesis | Aminolevulinate dehydratase, a homo-octameric enzyme, catalyzes the conversion of 5-aminolevulinate to porphobilinogen, the second step in heme biosynthesis; enzymatic activity is zinc-dependent; localizes to the cytoplasm and nucleus |
| HGH1 | HmG1/2 Homolog | Nonessential protein of unknown function; predicted to be involved in ribosome biogenesis; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm; similar to mammalian BRP16 (Brain protein 16) |
| HMG1 | 3-Hydroxy-3-MethylGlutaryl-coenzyme a reductase | One of two isozymes of HMG-CoA reductase that catalyzes the conversion of HMG-CoA to mevalonate, which is a rate-limiting step in sterol biosynthesis; localizes to the nuclear envelope; overproduction induces the formation of karmellae |
| HYP2 | HYPusine-containing protein | Translation elongation factor eIF-5A, previously thought to function in translation initiation; similar to and functionally redundant with Anb1p; structural homolog of bacterial EF-P; undergoes an essential hypusination modification |
| IDI1 | Isopentenyl Diphosphate Isomerase | Isopentenyl diphosphate:dimethylallyl diphosphate isomerase (IPP isomerase), catalyzes an essential activation step in the isoprenoid biosynthetic pathway; required for viability |

*continued on the next page*

**Genes in a putative sterol-metabolism and electron-transport cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| MTC7 | Maintenance of Telomere Capping | Predicted metabolic role based on network analysis derived from ChIP experiments, a large-scale deletion study and localization of transcription factor binding sites; null mutant is sensitive to temperature oscillation in a cdc13-1 mutant |
| MVD1 | MeValonate pyrophosphate Decarboxylase | Mevalonate pyrophosphate decarboxylase, essential enzyme involved in the biosynthesis of isoprenoids and sterols, including ergosterol; acts as a homodimer |
| PET100 | PETite colonies | Chaperone that specifically facilitates the assembly of cytochrome c oxidase, integral to the mitochondrial inner membrane; interacts with a subcomplex of subunits VII, VIIa, and VIII (Cox7p, Cox9p, and Cox8p) but not with the holoenzyme |
| PEX10 | PEroXin | Peroxisomal membrane E3 ubiquitin ligase required for for Ubc4p-dependent Pex5p ubiquitination and peroxisomal matrix protein import; contains zinc-binding RING domain; mutations in human homolog cause various peroxisomal disorders |
| QCR6 | ubiQuinol-cytochrome C oxidoReductase | Subunit 6 of the ubiquinol cytochrome-c reductase complex, which is a component of the mitochondrial inner membrane electron transport chain; highly acidic protein; required for maturation of cytochrome c1 |
| RIP1 | Rieske Iron-sulfur Protein | Ubiquinol-cytochrome-c reductase, a Rieske iron-sulfur protein of the mitochondrial cytochrome bc1 complex; transfers electrons from ubiquinol to cytochrome c1 during respiration |
| RNA14 | RNA synthesis | Cleavage and polyadenylation factor I (CF I) component involved in cleavage and polyadenylation of mRNA 3' ends; bridges interaction between Rna15p and Hrp1p in the CF I complex |
| SAM2 | S-AdenosylMethionine requiring | S-adenosylmethionine synthetase, catalyzes transfer of the adenosyl group of ATP to the sulfur atom of methionine; one of two differentially regulated isozymes (Sam1p and Sam2p) |
| SCM4 | Suppressor of Cdc4 Mutation | Potential regulatory effector of CDC4 function, suppresses a temperature-sensitive allele of CDC4, tripartite protein structure in which a charged region separates two uncharged domains, not essential for mitosis or meiosis |
| TIF6 | Translation Initiation Factor | Constituent of 66S pre-ribosomal particles, has similarity to human translation initiation factor 6 (eIF6); may be involved in the biogenesis and or stability of 60S ribosomal subunits |
| YHB1 | Yeast HemogloBin-like protein | Nitric oxide oxidoreductase, flavohemoglobin involved in nitric oxide detoxification; plays a role in the oxidative and nitrosative stress responses |
| YDL086W | N/A | Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies; YDL086W is not an essential gene |
| YDR340W | N/A | Putative protein of unknown function |
| YGR235C | N/A | Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies |
| YLR050C | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the endoplasmic reticulum; YLR050C is not an essential gene |

*continued on the next page*

**Genes in a putative sterol-metabolism and electron-transport cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| YLR101C | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified, essential ORF ERG27/YLR100W |

Table A.2: **Genes in a putative amino-acid- and protein-biosynthesis cluster formed by JFuzzy-K from data of** Brem *et al.* **(2002).**

| Name | Short description | Full description |
|---|---|---|
| AAT2 | Aspartate AminoTransferase | Cytosolic aspartate aminotransferase, involved in nitrogen metabolism; localizes to peroxisomes in oleate-grown cells |
| ADE3 | ADEnine requiring | Cytoplasmic trifunctional enzyme C1-tetrahydrofolate synthase, involved in single carbon metabolism and required for biosynthesis of purines, thymidylate, methionine, and histidine; null mutation causes auxotrophy for adenine and histidine |
| ALD5 | ALdehyde Dehydrogenase | Mitochondrial aldehyde dehydrogenase, involved in regulation or biosynthesis of electron transport chain components and acetate formation; activated by K+; utilizes NADP+ as the preferred coenzyme; constitutively expressed |
| ARO1 | AROmatic amino acid requiring | Pentafunctional arom protein, catalyzes steps 2 through 6 in the biosynthesis of chorismate, which is a precursor to aromatic amino acids |
| ARO4 | AROmatic amino acid requiring | 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase, catalyzes the first step in aromatic amino acid biosynthesis and is feedback-inhibited by tyrosine or high concentrations of phenylalanine or tryptophan |
| BAP2 | Branched-chain Amino acid Permease | High-affinity leucine permease, functions as a branched-chain amino acid permease involved in the uptake of leucine, isoleucine and valine; contains 12 predicted transmembrane domains |
| BAT1 | Branched-chain Amino acid Transaminase | Mitochondrial branched-chain amino acid aminotransferase, homolog of murine ECA39; highly expressed during logarithmic phase and repressed during stationary phase |
| CAB1 | Coenzyme A Biosynthesis | Pantothenate kinase (ATP:D-pantothenate 4'-phosphotransferase, EC 2.7.1.33); catalyzes the first committed step in the universal biosynthetic pathway for synthesis of coenzyme A (CoA) |
| CIT2 | CITrate synthase | Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle; expression is controlled by Rtg1p and Rtg2p transcription factors |
| DIC1 | DIcarboxylate Carrier | Mitochondrial dicarboxylate carrier, integral membrane protein, catalyzes a dicarboxylate-phosphate exchange across the inner mitochondrial membrane, transports cytoplasmic dicarboxylates into the mitochondrial matrix |
| ECM17 | METhionine requiring | Sulfite reductase beta subunit, involved in amino acid biosynthesis, transcription repressed by methionine |
| FRE6 | Ferric REductase | Putative ferric reductase with similarity to Fre2p; expression induced by low iron levels |
| GCD14 | General Control Nonderepressible | Subunit of tRNA (1-methyladenosine) methyltransferase, with Gcd10p, required for the modification of the adenine at position 58 in tRNAs, especially tRNAi-Met; first identified as a negative regulator of GCN4 expression |

*continued on the next page*

**Genes in a putative amino-acid- and protein-biosynthesis cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| GCN1 | General Control Non-derepressible | Positive regulator of the Gcn2p kinase activity, forms a complex with Gcn20p; proposed to stimulate Gcn2p activation by an uncharged tRNA |
| GCN4 | General Control Non-derepressible | Basic leucine zipper (bZIP) transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels |
| GDH1 | Glutamate DeHydrogenase | NADP(+)-dependent glutamate dehydrogenase, synthesizes glutamate from ammonia and alpha-ketoglutarate; rate of alpha-ketoglutarate utilization differs from Gdh3p; expression regulated by nitrogen and carbon sources |
| GDH3 | Glutamate DeHydrogenase | NADP(+)-dependent glutamate dehydrogenase, synthesizes glutamate from ammonia and alpha-ketoglutarate; rate of alpha-ketoglutarate utilization differs from Gdh1p; expression regulated by nitrogen and carbon sources |
| GTR2 | GTp binding protein Resemblance | Putative GTP binding protein that negatively regulates Ran/Tc4 GTPase cycle; activates transcription; subunit of EGO and GSE complexes; required for sorting of Gap1p; localizes to cytoplasm and to chromatin; homolog of human RagC and RagD |
| HAC1 | Homologous to Atf/Creb1 | Basic leucine zipper (bZIP) transcription factor (ATF/CREB1 homolog) that regulates the unfolded protein response, via UPRE binding, and membrane biogenesis; ER stress-induced splicing pathway facilitates efficient Hac1p synthesis |
| HIP1 | HIstidine Permease | High-affinity histidine permease, also involved in the transport of manganese ions |
| HIS5 | HIStidine requiring | Histidinol-phosphate aminotransferase, catalyzes the seventh step in histidine biosynthesis; responsive to general control of amino acid biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts |
| HOM2 | HOMoserine requiring | Aspartic beta semi-aldehyde dehydrogenase, catalyzes the second step in the common pathway for methionine and threonine biosynthesis; expression regulated by Gcn4p and the general control of amino acid synthesis |
| HRB1 | Hypothetical RNA-Binding protein | Poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm; similar to Gbp2p and Npl3p |
| ILV2 | IsoLeucine-plus-Valine requiring | Acetolactate synthase, catalyses the first common step in isoleucine and valine biosynthesis and is the target of several classes of inhibitors, localizes to the mitochondria; expression of the gene is under general amino acid control |
| ILV3 | IsoLeucine-plus-Valine requiring | Dihydroxyacid dehydratase, catalyzes third step in the common pathway leading to biosynthesis of branched-chain amino acids |
| ILV5 | IsoLeucine-plus-Valine requiring | Acetohydroxyacid reductoisomerase, mitochondrial protein involved in branched-chain amino acid biosynthesis, also required for maintenance of wild-type mitochondrial DNA and found in mitochondrial nucleoids |
| ILV6 | IsoLeucine-plus-Valine requiring | Regulatory subunit of acetolactate synthase, which catalyzes the first step of branched-chain amino acid biosynthesis; enhances activity of the Ilv2p catalytic subunit, localizes to mitochondria |

*continued on the next page*

**Genes in a putative amino-acid- and protein-biosynthesis cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| ISU1 | Iron-Sulfur cluster nifU-like protein | Conserved protein of the mitochondrial matrix, performs a scaffolding function during assembly of iron-sulfur clusters, interacts physically and functionally with yeast frataxin (Yfh1p); isu1 isu2 double mutant is inviable |
| ISU2 | Iron-Sulfur cluster nifU-like protein | Conserved protein of the mitochondrial matrix, required for synthesis of mitochondrial and cytosolic iron-sulfur proteins, performs a scaffolding function in mitochondria during Fe/S cluster assembly; isu1 isu2 double mutant is inviable |
| LEU1 | LEUcine biosynthesis | Isopropylmalate isomerase, catalyzes the second step in the leucine biosynthesis pathway |
| LEU4 | LEUcine biosynthesis | Alpha-isopropylmalate synthase (2-isopropylmalate synthase); the main isozyme responsible for the first step in the leucine biosynthesis pathway |
| LEU9 | LEUcine biosynthesis | Alpha-isopropylmalate synthase II (2-isopropylmalate synthase), catalyzes the first step in the leucine biosynthesis pathway; the minor isozyme, responsible for the residual alpha-IPMS activity detected in a leu4 null mutant |
| LYS4 | LYSine requiring | Homoaconitase, catalyzes the conversion of homocitrate to homoisocitrate, which is a step in the lysine biosynthesis pathway |
| MAE1 | MAlic Enzyme | Mitochondrial malic enzyme, catalyzes the oxidative decarboxylation of malate to pyruvate, which is a key intermediate in sugar metabolism and a precursor for synthesis of several amino acids |
| MCT1 | Malonyl-CoA:ACP Transferase | Predicted malonyl-CoA:ACP transferase, putative component of a type-II mitochondrial fatty acid synthase that produces intermediates for phospholipid remodeling |
| MET1 | METhionine requiring | S-adenosyl-L-methionine uroporphyrinogen III transmethylase, involved in the biosynthesis of siroheme, a prosthetic group used by sulfite reductase; required for sulfate assimilation and methionine biosynthesis |
| MET10 | METhionine requiring | Subunit alpha of assimilatory sulfite reductase, which converts sulfite into sulfide |
| MET12 | METhionine requiring | Protein with methylenetetrahydrofolate reductase (MTHFR) activity in vitro; null mutant has no phenotype and is prototrophic for methionine; MET13 encodes major isozyme of MTHFR |
| MET13 | METhionine requiring | Major isozyme of methylenetetrahydrofolate reductase, catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate in the methionine biosynthesis pathway |
| MET14 | METhionine requiring | Adenylylsulfate kinase, required for sulfate assimilation and involved in methionine metabolism |
| MET16 | METhionine requiring | 3'-phosphoadenylsulfate reductase, reduces 3'-phosphoadenylyl sulfate to adenosine-3',5'-bisphosphate and free sulfite using reduced thioredoxin as cosubstrate, involved in sulfate assimilation and methionine metabolism |
| MET17 | METhionine requiring | Methionine and cysteine synthase (O-acetyl homoserine-O-acetyl serine sulfhydrylase), required for sulfur amino acid synthesis |
| MET18 | METhionine requiring | DNA repair and TFIIH regulator, required for both nucleotide excision repair (NER) and RNA polymerase II (RNAP II) transcription; involved in telomere maintenance |

*continued on the next page*

**Genes in a putative amino-acid- and protein-biosynthesis cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| MET22 | METhionine requiring | Bisphosphate-3'-nucleotidase, involved in salt tolerance and methionine biogenesis; dephosphorylates 3'-phosphoadenosine-5'-phosphate and 3'-phosphoadenosine-5'-phosphosulfate, intermediates of the sulfate assimilation pathway |
| MUP3 | Methionine UPtake | Low affinity methionine permease, similar to Mup1p |
| NCE103 | NonClassical Export | Carbonic anhydrase; poorly transcribed under aerobic conditions and at an undetectable level under anaerobic conditions; involved in non-classical protein export pathway |
| NFS1 | NiFS-like | Cysteine desulfurase involved in iron-sulfur cluster (Fe/S) biogenesis; required for the post-transcriptional thio-modification of mitochondrial and cytoplasmic tRNAs; essential protein located predominantly in mitochondria |
| OAC1 | OxaloAcetate Carrier | Mitochondrial inner membrane transporter, transports oxaloacetate, sulfate, thiosulfate, and isopropylmalate; member of the mitochondrial carrier family |
| PCL7 | Pho85 CycLin | Pho85p cyclin of the Pho80p subfamily, forms a functional kinase complex with Pho85p which phosphorylates Mmr1p and is regulated by Pho81p; involved in glycogen metabolism, expression is cell-cycle regulated |
| PYC1 | PYruvate Carboxylase | Pyruvate carboxylase isoform, cytoplasmic enzyme that converts pyruvate to oxaloacetate; highly similar to isoform Pyc2p but differentially regulated; mutations in the human homolog are associated with lactic acidosis |
| RGT2 | Restores Glucose Transport | Plasma membrane glucose receptor, highly similar to Snf3p; both Rgt2p and Snf3p serve as transmembrane glucose sensors generating an intracellular signal that induces expression of glucose transporter (HXT) genes |
| RIB1 | RIBoflavin biosynthesis | GTP cyclohydrolase II; catalyzes the first step of the riboflavin biosynthesis pathway |
| RIB3 | RIBoflavin biosynthesis | 3,4-dihydroxy-2-butanone-4-phosphate synthase (DHBP synthase), required for riboflavin biosynthesis from ribulose-5-phosphate, also has an unrelated function in mitochondrial respiration |
| RPA14 | RNA Polymerase A | RNA polymerase I subunit A14 |
| RTG2 | ReTroGrade regulation | Sensor of mitochondrial dysfunction; regulates the subcellular location of Rtg1p and Rtg3p, transcriptional activators of the retrograde (RTG) and TOR pathways; Rtg2p is inhibited by the phosphorylated form of Mks1p |
| RTG3 | ReTroGrade regulation | Basic helix-loop-helix-leucine zipper (bHLH/Zip) transcription factor that forms a complex with another bHLH/Zip protein, Rtg1p, to activate the retrograde (RTG) and TOR pathways |
| SNQ2 | Sensitivity to 4-NitroQuinoline-N-oxide | Plasma membrane ATP-binding cassette (ABC) transporter, multidrug transporter involved in multidrug resistance and resistance to singlet oxygen species |
| SPE1 | SPErmidine auxotroph | Ornithine decarboxylase, catalyzes the first step in polyamine biosynthesis; degraded in a proteasome-dependent manner in the presence of excess polyamines |

*continued on the next page*

**Genes in a putative amino-acid- and protein-biosynthesis cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| STB4 | Sin Three Binding protein | Protein that binds Sin3p in a two-hybrid assay; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters |
| TAH11 | Topo-A Hypersensitive | DNA replication licensing factor, required for pre-replication complex assembly |
| TAH18 | Top1T722A mutant Hypersensitive | Protein of unknown function that plays a pro-death role in response to oxidative stress; highly conserved across species and similar to human protein NDOR1; allele is synthetically lethal with the pol3-13 allele of DNA polymerase delta |
| TFB3 | Transcription initiation Factor IIB | Subunit of TFIIH and nucleotide excision repair factor 3 complexes, involved in transcription initiation, required for nucleotide excision repair; ring finger protein similar to mammalian CAK and TFIIH subunit |
| THR1 | THReonine requiring | Homoserine kinase, conserved protein required for threonine biosynthesis; expression is regulated by the GCN4-mediated general amino acid control pathway |
| TRP4 | TRyPtophan requiring | Anthranilate phosphoribosyl transferase of the tryptophan biosynthetic pathway, catalyzes the phosphoribosylation of anthranilate, subject to the general control system of amino acid biosynthesis |
| XDJ1 | Putative chaperone | Putative chaperone, homolog of E. coli DnaJ, closely related to Ydj1p; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies |
| YPS7 | YaPSin | Putative GPI-anchored aspartic protease, located in the cytoplasm and endoplasmic reticulum |
| YRR1 | Yeast Reveromycin-A Resistant | Zn2-Cys6 zinc-finger transcription factor that activates genes involved in multidrug resistance; paralog of Yrm1p, acting on an overlapping set of target genes |
| YBR012C | N/A | Dubious open reading frame, unlikely to encode a functional protein; expression induced by iron-regulated transcriptional activator Aft2p |
| YDR476C | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the endoplasmic reticulum; YDR476C is not an essential gene |
| YFL032W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene HAC1/YFL031W; YFL032W is not an essential gene |
| YGL010W | N/A | Putative protein of unknown function; YGL010W is not an essential gene |
| YGL114W | N/A | Putative protein of unknown function; predicted member of the oligopeptide transporter (OPT) family of membrane transporters |
| YGL138C | N/A | Putative protein of unknown function; has no significant sequence similarity to any known protein |
| YGR146C | N/A | Putative protein of unknown function; induced by iron homeostasis transcription factor Aft2p; multicopy suppressor of a temperature sensitive hsf1 mutant; induced by treatment with 8-methoxypsoralen and UVA irradiation |

*continued on the next page*

**Genes in a putative amino-acid- and protein-biosynthesis cluster** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| YHR122W | N/A | Protein of unknown function required for establishment of sister chromatid cohesion; synthetically lethal with RFC5, an RF-C subunit that links replication to cohesion establishment; YHR122W is an essential gene |
| YHR162W | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the mitochondrion |
| YJL171C | N/A | GPI-anchored cell wall protein of unknown function; induced in response to cell wall damaging agents and by mutations in genes involved in cell wall biogenesis; sequence similarity to YBR162C/TOS1, a covalently bound cell wall protein |
| YKL030W | N/A | Dubious open reading frame, unlikely to encode a protein; not conserved in closely related Saccharomyces species; partially overlaps the verified gene MAE1 |
| YLR149C | N/A | Putative protein of unknown function; overexpression causes a cell cycle delay or arrest; null mutation results in a decrease in plasma membrane electron transport; YLR149C is not an essential gene |
| YLR179C | N/A | Protein of unknown function with similarity to Tfs1p; transcription is activated by paralogous proteins Yrm1p and Yrr1p along with proteins involved in multidrug resistance; GFP-tagged protein localizes to the cytoplasm and nucleus |
| YMR321C | N/A | Putative protein of unknown function; proposed to be a palmitoylated membrane protein |
| YNL276C | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; overlaps the verified gene MET2/YNL277W |
| YOR225W | N/A | Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data |
| YOR292C | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the vacuole; YOR292C is not an essential gene |

Table A.3: **Genes in putative amino-acid biosynthesis and -metabolism cluster 6 formed by JFuzzy-K from data of** Brem *et al.* **(2002).**

| Name | Short description | Full description |
|---|---|---|
| AAT1 | Aspartate AminoTransferase | Mitochondrial aspartate aminotransferase, catalyzes the conversion of oxaloacetate to aspartate in aspartate and asparagine biosynthesis |
| AAT2 | Aspartate AminoTransferase | Cytosolic aspartate aminotransferase, involved in nitrogen metabolism; localizes to peroxisomes in oleate-grown cells |
| ADE3 | ADEnine requiring | Cytoplasmic trifunctional enzyme C1-tetrahydrofolate synthase, involved in single carbon metabolism and required for biosynthesis of purines, thymidylate, methionine, and histidine; null mutation causes auxotrophy for adenine and histidine |
| ADE4 | ADEnine requiring | Phosphoribosylpyrophosphate amidotransferase (PRPPAT; amidophosphoribosyltransferase), catalyzes first step of the 'de novo' purine nucleotide biosynthetic pathway |
| ADE6 | ADEnine requiring | Formylglycinamidine-ribonucleotide (FGAM)-synthetase, catalyzes a step in the 'de novo' purine nucleotide biosynthetic pathway |
| ADE8 | ADEnine requiring | Phosphoribosyl-glycinamide transformylase, catalyzes a step in the 'de novo' purine nucleotide biosynthetic pathway |
| ADH5 | Alcohol DeHydrogenase | Alcohol dehydrogenase isoenzyme V; involved in ethanol production |
| ALD5 | ALdehyde Dehydrogenase | Mitochondrial aldehyde dehydrogenase, involved in regulation or biosynthesis of electron transport chain components and acetate formation; activated by K+; utilizes NADP+ as the preferred coenzyme; constitutively expressed |
| ALG2 | ALG2 | Mannosyltransferase that catalyzes two consecutive steps in the N-linked glycosylation pathway; alg2 mutants exhibit temperature-sensitive growth and abnormal accumulation of the lipid-linked oligosaccharide Man2GlcNAc2-PP-Dol |
| AMD2 | AMiDase | Putative amidase |
| APA2 | AP4A phosphorylase | Diadenosine 5',5"-P1,P4-tetraphosphate phosphorylase II (AP4A phosphorylase), involved in catabolism of bis(5'-nucleosidyl) tetraphosphates; has similarity to Apa1p |
| ARG2 | ARGinine requiring | Acetylglutamate synthase (glutamate N-acetyltransferase), mitochondrial enzyme that catalyzes the first step in the biosynthesis of the arginine precursor ornithine; forms a complex with Arg5,6p |
| ARG4 | ARGinine requiring | Argininosuccinate lyase, catalyzes the final step in the arginine biosynthesis pathway |
| ARG5,6 | ARGinine requiring | Protein that is processed in the mitochondrion to yield acetylglutamate kinase and N-acetyl-gamma-glutamyl-phosphate reductase, which catalyze the 2nd and 3rd steps in arginine biosynthesis; enzymes form a complex with Arg2p |
| ARG80 | ARGinine requiring | Transcription factor involved in regulation of arginine-responsive genes; acts with Arg81p and Arg82p |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| ARG81 | ARGinine requiring | Zinc-finger transcription factor of the Zn(2)-Cys(6) binuclear cluster domain type, involved in the regulation of arginine-responsive genes; acts with Arg80p and Arg82p |
| ARG82 | ARGinine requiring | Inositol polyphosphate multikinase (IPMK), sequentially phosphorylates Ins(1,4,5)P3 to form Ins(1,3,4,5,6)P5; also has diphosphoinositol polyphosphate synthase activity; regulates arginine-, phosphate-, and nitrogen-responsive genes |
| ARG8 | ARGinine requiring | Acetylornithine aminotransferase, catalyzes the fourth step in the biosynthesis of the arginine precursor ornithine |
| ARG80 | ARGinine requiring | Transcription factor involved in regulation of arginine-responsive genes; acts with Arg81p and Arg82p |
| ARG81 | ARGinine requiring | Zinc-finger transcription factor of the Zn(2)-Cys(6) binuclear cluster domain type, involved in the regulation of arginine-responsive genes; acts with Arg80p and Arg82p |
| ARG82 | ARGinine requiring | Inositol polyphosphate multikinase (IPMK), sequentially phosphorylates Ins(1,4,5)P3 to form Ins(1,3,4,5,6)P5; also has diphosphoinositol polyphosphate synthase activity; regulates arginine-, phosphate-, and nitrogen-responsive genes |
| ARO10 | AROmatic amino acid requiring | Phenylpyruvate decarboxylase, catalyzes decarboxylation of phenylpyruvate to phenylacetaldehyde, which is the first specific step in the Ehrlich pathway |
| ARO1 | AROmatic amino acid requiring | Pentafunctional arom protein, catalyzes steps 2 through 6 in the biosynthesis of chorismate, which is a precursor to aromatic amino acids |
| ARO2 | AROmatic amino acid requiring | Bifunctional chorismate synthase and flavin reductase, catalyzes the conversion of 5-enolpyruvylshikimate 3-phosphate (EPSP) to form chorismate, which is a precursor to aromatic amino acids |
| ARO3 | AROmatic amino acid requiring | 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase, catalyzes the first step in aromatic amino acid biosynthesis and is feedback-inhibited by phenylalanine or high concentration of tyrosine or tryptophan |
| ARO4 | AROmatic amino acid requiring | 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase, catalyzes the first step in aromatic amino acid biosynthesis and is feedback-inhibited by tyrosine or high concentrations of phenylalanine or tryptophan |
| ARO80 | AROmatic amino acid requiring | Zinc finger transcriptional activator of the Zn2Cys6 family; activates transcription of aromatic amino acid catabolic genes in the presence of aromatic amino acids |
| ARO8 | AROmatic amino acid requiring | Aromatic aminotransferase I, expression is regulated by general control of amino acid biosynthesis |
| ARP10 | Actin-Related Protein | Component of the dynactin complex, localized to the pointed end of the Arp1p filament; may regulate membrane association of the complex |
| ASK10 | Activator of SKn7 | Component of the RNA polymerase II holoenzyme, phosphorylated in response to oxidative stress; has a role in destruction of Ssn8p, which relieves repression of stress-response genes |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| ASN1 | ASparagiNe requiring | Asparagine synthetase, isozyme of Asn2p; catalyzes the synthesis of L-asparagine from L-aspartate in the asparagine biosynthetic pathway |
| ASN2 | ASparagiNe requiring | Asparagine synthetase, isozyme of Asn1p; catalyzes the synthesis of L-asparagine from L-aspartate in the asparagine biosynthetic pathway |
| ATR1 | AminoTriazole Resistance | Multidrug efflux pump of the major facilitator superfamily, required for resistance to aminotriazole and 4-nitroquinoline-N-oxide |
| ATS1 | Alpha Tubulin Suppressor | Protein required, with Elongator complex, Kti11p, and Kti12p, for modification of wobble nucleosides in tRNA; has a potential role in regulatory interactions between microtubules and the cell cycle |
| BAP3 | Branched-chain Amino acid Permease | Amino acid permease involved in the uptake of cysteine, leucine, isoleucine and valine |
| BIG1 | Bad In Glucose | Integral membrane protein of the endoplasmic reticulum, required for normal content of cell wall beta-1,6-glucan |
| BOP2 | Bypass Of Pam1 | Protein of unknown function |
| BRO1 | BCK1-like Resistance to Osmotic shock | Cytoplasmic class E vacuolar protein sorting (VPS) factor that coordinates deubiquitination in the multivesicular body (MVB) pathway by recruiting Doa4p to endosomes |
| CAF16 | CCR4 Associated Factor | Part of the evolutionarily-conserved CCR4-NOT transcriptional regulatory complex involved in controlling mRNA initiation, elongation, and degradation; putative ABC ATPase; interacts with Ssn2p, Ssn3p, and Ssn8p |
| CAN1 | CANavanine resistance | Plasma membrane arginine permease, requires phosphatidyl ethanolamine (PE) for localization, exclusively associated with lipid rafts; mutation confers canavanine resistance |
| CDC54 | MiniChromosome Maintenance | Essential helicase component of heterohexameric MCM2-7 complexes which bind pre-replication complexes on DNA and melt the DNA prior to replication; accumulates in the nucleus in G1; homolog of S. pombe Cdc21p |
| CIT2 | CITrate synthase | Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle; expression is controlled by Rtg1p and Rtg2p transcription factors |
| CLG1 | Cyclin-Like Gene | Cyclin-like protein that interacts with Pho85p; has sequence similarity to G1 cyclins PCL1 and PCL2 |
| COT1 | CObalt Toxicity | Vacuolar transporter that mediates zinc transport into the vacuole; overexpression confers resistance to cobalt and rhodium |
| CPA2 | Carbamyl Phosphate synthetase A | Large subunit of carbamoyl phosphate synthetase, which catalyzes a step in the synthesis of citrulline, an arginine precursor |
| CTF13 | Chromosome Transmission Fidelity | Subunit of the CBF3 complex, which binds to the CDE III element of centromeres, bending the DNA upon binding, and may be involved in sister chromatid cohesion during mitosis |
| DBF20 | DumbBell Forming | Ser/Thr kinase involved in late nuclear division, one of the mitotic exit network (MEN) proteins; necessary for the execution of cytokinesis |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| DBF2 | DumbBell Forming | Ser/Thr kinase involved in transcription and stress response; functions as part of a network of genes in exit from mitosis; localization is cell cycle regulated; activated by Cdc15p during the exit from mitosis |
| DBF20 | DumbBell Forming | Ser/Thr kinase involved in late nuclear division, one of the mitotic exit network (MEN) proteins; necessary for the execution of cytokinesis |
| DIC1 | DIcarboxylate Carrier | Mitochondrial dicarboxylate carrier, integral membrane protein, catalyzes a dicarboxylate-phosphate exchange across the inner mitochondrial membrane, transports cytoplasmic dicarboxylates into the mitochondrial matrix |
| ECM17 | METhionine requiring | Sulfite reductase beta subunit, involved in amino acid biosynthesis, transcription repressed by methionine |
| ECM40 | ARGinine requiring | Mitochondrial ornithine acetyltransferase, catalyzes the fifth step in arginine biosynthesis; also possesses acetylglutamate synthase activity, regenerates acetylglutamate while forming ornithine |
| ERV14 | ER Vesicle | Protein localized to COPII-coated vesicles, involved in vesicle formation and incorporation of specific secretory cargo; required for the delivery of bud-site selection protein Axl2p to cell surface; related to Drosophila cornichon |
| ERV15 | ER Vesicle Protein | Protein involved in export of proteins from the endoplasmic reticulum, has similarity to Erv14p |
| ERV1 | Essential for Respiration and Viability | Flavin-linked sulfhydryl oxidase of the mitochondrial intermembrane space (IMS), oxidizes Mia40p as part of a disulfide relay system that promotes IMS retention of imported proteins; ortholog of human hepatopoietin (ALR) |
| ERV14 | ER Vesicle | Protein localized to COPII-coated vesicles, involved in vesicle formation and incorporation of specific secretory cargo; required for the delivery of bud-site selection protein Axl2p to cell surface; related to Drosophila cornichon |
| ESBP6 | MCH3 | Protein with similarity to monocarboxylate permeases, appears not to be involved in transport of monocarboxylates such as lactate, pyruvate or acetate across the plasma membrane |
| FOL2 | FOLic acid synthesis | GTP-cyclohydrolase I, catalyzes the first step in the folic acid biosynthetic pathway |
| FRE6 | Ferric REductase | Putative ferric reductase with similarity to Fre2p; expression induced by low iron levels |
| GAT1 | GATA-1-type zinc finger | Transcriptional activator of genes involved in nitrogen catabolite repression; contains a GATA-1-type zinc finger DNA-binding motif; activity and localization regulated by nitrogen limitation and Ure2p |
| GDH2 | Glutamate DeHydrogenase | NAD(+)-dependent glutamate dehydrogenase, degrades glutamate to ammonia and alpha-ketoglutarate; expression sensitive to nitrogen catabolite repression and intracellular ammonia levels |
| GLN3 | GLutamiNe metabolism | Transcriptional activator of genes regulated by nitrogen catabolite repression (NCR), localization and activity regulated by quality of nitrogen source |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| GLT1 | GLuTamate synthase | NAD(+)-dependent glutamate synthase (GOGAT), synthesizes glutamate from glutamine and alpha-ketoglutarate; with Gln1p, forms the secondary pathway for glutamate biosynthesis from ammonia; expression regulated by nitrogen source |
| GLY1 | GLYcine requiring | Threonine aldolase, catalyzes the cleavage of L-allo-threonine and L-threonine to glycine; involved in glycine biosynthesis |
| GPD2 | Glycerol-3-Phosphate Dehydrogenase | NAD-dependent glycerol 3-phosphate dehydrogenase, homolog of Gpd1p, expression is controlled by an oxygen-independent signaling pathway required to regulate metabolism under anoxic conditions; located in cytosol and mitochondria |
| GTR2 | GTp binding protein Resemblance | Putative GTP binding protein that negatively regulates Ran/Tc4 GTPase cycle; activates transcription; subunit of EGO and GSE complexes; required for sorting of Gap1p; localizes to cytoplasm and to chromatin; homolog of human RagC and RagD |
| HIP1 | HIstidine Permease | High-affinity histidine permease, also involved in the transport of manganese ions |
| HIS1 | HIStidine requiring | ATP phosphoribosyltransferase, a hexameric enzyme, catalyzes the first step in histidine biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts; transcription is regulated by general amino acid control |
| HIS2 | HIStidine requiring | Histidinolphosphatase, catalyzes the eighth step in histidine biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts; transcription is regulated by general amino acid control |
| HIS3 | HIStidine requiring | Imidazoleglycerol-phosphate dehydratase, catalyzes the sixth step in histidine biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts; transcription is regulated by general amino acid control via Gcn4p |
| HIS4 | HIStidine requiring | Multifunctional enzyme containing phosphoribosyl-ATP pyrophosphatase, phosphoribosyl-AMP cyclohydrolase, and histidinol dehydrogenase activities; catalyzes the second, third, ninth and tenth steps in histidine biosynthesis |
| HIS5 | HIStidine requiring | Histidinol-phosphate aminotransferase, catalyzes the seventh step in histidine biosynthesis; responsive to general control of amino acid biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts |
| HIS7 | HIStidine requiring | Imidazole glycerol phosphate synthase (glutamine amidotransferase:cyclase), catalyzes the fifth and sixth steps of histidine biosynthesis and also produces 5-aminoimidazole-4-carboxamide ribotide (AICAR), a purine precursor |
| HOM2 | HOMoserine requiring | Aspartic beta semi-aldehyde dehydrogenase, catalyzes the second step in the common pathway for methionine and threonine biosynthesis; expression regulated by Gcn4p and the general control of amino acid synthesis |
| HOM3 | HOMoserine requiring | Aspartate kinase (L-aspartate 4-P-transferase); cytoplasmic enzyme that catalyzes the first step in the common pathway for methionine and threonine biosynthesis; expression regulated by Gcn4p and the general control of amino acid synthesis |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| HRB1 | Hypothetical RNA-Binding protein | Poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm; similar to Gbp2p and Npl3p |
| IDP1 | Isocitrate Dehydrogenase, NADP-specific | Mitochondrial NADP-specific isocitrate dehydrogenase, catalyzes the oxidation of isocitrate to alpha-ketoglutarate; not required for mitochondrial respiration and may function to divert alpha-ketoglutarate to biosynthetic processes |
| ILV1 | IsoLeucine-plus-Valine requiring | Threonine deaminase, catalyzes the first step in isoleucine biosynthesis; expression is under general amino acid control; ILV1 locus exhibits highly positioned nucleosomes whose organization is independent of known ILV1 regulation |
| ILV3 | IsoLeucine-plus-Valine requiring | Dihydroxyacid dehydratase, catalyzes third step in the common pathway leading to biosynthesis of branched-chain amino acids |
| ILV5 | IsoLeucine-plus-Valine requiring | Acetohydroxyacid reductoisomerase, mitochondrial protein involved in branched-chain amino acid biosynthesis, also required for maintenance of wild-type mitochondrial DNA and found in mitochondrial nucleoids |
| ILV6 | IsoLeucine-plus-Valine requiring | Regulatory subunit of acetolactate synthase, which catalyzes the first step of branched-chain amino acid biosynthesis; enhances activity of the Ilv2p catalytic subunit, localizes to mitochondria |
| ISA1 | Iron Sulfur Assembly | Mitochondrial matrix protein involved in biogenesis of the iron-sulfur (Fe/S) cluster of Fe/S proteins, isa1 deletion causes loss of mitochondrial DNA and respiratory deficiency; depletion reduces growth on nonfermentable carbon sources |
| ISU1 | Iron-Sulfur cluster nifU-like protein | Conserved protein of the mitochondrial matrix, performs a scaffolding function during assembly of iron-sulfur clusters, interacts physically and functionally with yeast frataxin (Yfh1p); isu1 isu2 double mutant is inviable |
| ISU2 | Iron-Sulfur cluster nifU-like protein | Conserved protein of the mitochondrial matrix, required for synthesis of mitochondrial and cytosolic iron-sulfur proteins, performs a scaffolding function in mitochondria during Fe/S cluster assembly; isu1 isu2 double mutant is inviable |
| KAR9 | KARyogamy | Karyogamy protein required for correct positioning of the mitotic spindle and for orienting cytoplasmic microtubules, localizes at the shmoo tip in mating cells and at the tip of the growing bud in small-budded cells through anaphase |
| KRS1 | Lysyl (K) tRNA Synthetase | Lysyl-tRNA synthetase |
| LAP3 | Leucine AminoPeptidases | Cysteine aminopeptidase with homocysteine-thiolactonase activity; protects cells against homocysteine toxicity; has bleomycin hydrolase activity in vitro; transcription is regulated by galactose via Gal4p; orthologous to human BLMH |
| LEU4 | LEUcine biosynthesis | Alpha-isopropylmalate synthase (2-isopropylmalate synthase); the main isozyme responsible for the first step in the leucine biosynthesis pathway |
| LIP5 | LIPoic acid | Protein involved in biosynthesis of the coenzyme lipoic acid, has similarity to E. coli lipoic acid synthase |
| LYP1 | Lysine permease | Lysine permease; one of three amino acid permeases (Alp1p, Can1p, Lyp1p) responsible for uptake of cationic amino acids |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| LYS12 | LYSine requiring | Homo-isocitrate dehydrogenase, an NAD-linked mitochondrial enzyme required for the fourth step in the biosynthesis of lysine, in which homo-isocitrate is oxidatively decarboxylated to alpha-ketoadipate |
| LYS14 | LYSine requiring | Transcriptional activator involved in regulation of genes of the lysine biosynthesis pathway; requires 2-aminoadipate semialdehyde as co-inducer |
| LYS1 | LYSine requiring | Saccharopine dehydrogenase (NAD+, L-lysine-forming), catalyzes the conversion of saccharopine to L-lysine, which is the final step in the lysine biosynthesis pathway |
| LYS12 | LYSine requiring | Homo-isocitrate dehydrogenase, an NAD-linked mitochondrial enzyme required for the fourth step in the biosynthesis of lysine, in which homo-isocitrate is oxidatively decarboxylated to alpha-ketoadipate |
| LYS14 | LYSine requiring | Transcriptional activator involved in regulation of genes of the lysine biosynthesis pathway; requires 2-aminoadipate semialdehyde as co-inducer |
| LYS20 | LYSine requiring | Homocitrate synthase isozyme, catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate to form homocitrate, which is the first step in the lysine biosynthesis pathway; highly similar to the other isozyme, Lys21p |
| LYS21 | LYSine requiring | Homocitrate synthase isozyme, catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate to form homocitrate, which is the first step in the lysine biosynthesis pathway; highly similar to the other isozyme, Lys20p |
| LYS4 | LYSine requiring | Homoaconitase, catalyzes the conversion of homocitrate to homoisocitrate, which is a step in the lysine biosynthesis pathway |
| LYS9 | LYSine requiring | Saccharopine dehydrogenase (NADP+, L-glutamate-forming); catalyzes the formation of saccharopine from alpha-aminoadipate 6-semialdehyde, the seventh step in lysine biosynthesis pathway; exhibits genetic and physical interactions with TRM112 |
| MAC1 | Metal binding ACtivator | Copper-sensing transcription factor involved in regulation of genes required for high affinity copper transport |
| MAL13 | MALtose fermentation | MAL-activator protein, part of complex locus MAL1; nonfunctional in genomic reference strain S288C |
| MCT1 | Malonyl-CoA:ACP Transferase | Predicted malonyl-CoA:ACP transferase, putative component of a type-II mitochondrial fatty acid synthase that produces intermediates for phospholipid remodeling |
| MET10 | METhionine requiring | Subunit alpha of assimilatory sulfite reductase, which converts sulfite into sulfide |
| MET12 | METhionine requiring | Protein with methylenetetrahydrofolate reductase (MTHFR) activity in vitro; null mutant has no phenotype and is prototrophic for methionine; MET13 encodes major isozyme of MTHFR |
| MET13 | METhionine requiring | Major isozyme of methylenetetrahydrofolate reductase, catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate in the methionine biosynthesis pathway |
| MET14 | METhionine requiring | Adenylylsulfate kinase, required for sulfate assimilation and involved in methionine metabolism |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| MET16 | METhionine requiring | 3'-phosphoadenylsulfate reductase, reduces 3'-phosphoadenylyl sulfate to adenosine-3',5'-bisphosphate and free sulfite using reduced thioredoxin as cosubstrate, involved in sulfate assimilation and methionine metabolism |
| MET17 | METhionine requiring | Methionine and cysteine synthase (O-acetyl homoserine-O-acetyl serine sulfhydrylase), required for sulfur amino acid synthesis |
| MET18 | METhionine requiring | DNA repair and TFIIH regulator, required for both nucleotide excision repair (NER) and RNA polymerase II (RNAP II) transcription; involved in telomere maintenance |
| MET1 | METhionine requiring | S-adenosyl-L-methionine uroporphyrinogen III transmethylase, involved in the biosynthesis of siroheme, a prosthetic group used by sulfite reductase; required for sulfate assimilation and methionine biosynthesis |
| MET10 | METhionine requiring | Subunit alpha of assimilatory sulfite reductase, which converts sulfite into sulfide |
| MET12 | METhionine requiring | Protein with methylenetetrahydrofolate reductase (MTHFR) activity in vitro; null mutant has no phenotype and is prototrophic for methionine; MET13 encodes major isozyme of MTHFR |
| MET13 | METhionine requiring | Major isozyme of methylenetetrahydrofolate reductase, catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate in the methionine biosynthesis pathway |
| MET14 | METhionine requiring | Adenylylsulfate kinase, required for sulfate assimilation and involved in methionine metabolism |
| MET16 | METhionine requiring | 3'-phosphoadenylsulfate reductase, reduces 3'-phosphoadenylyl sulfate to adenosine-3',5'-bisphosphate and free sulfite using reduced thioredoxin as cosubstrate, involved in sulfate assimilation and methionine metabolism |
| MET17 | METhionine requiring | Methionine and cysteine synthase (O-acetyl homoserine-O-acetyl serine sulfhydrylase), required for sulfur amino acid synthesis |
| MET18 | METhionine requiring | DNA repair and TFIIH regulator, required for both nucleotide excision repair (NER) and RNA polymerase II (RNAP II) transcription; involved in telomere maintenance |
| MET22 | METhionine requiring | Bisphosphate-3'-nucleotidase, involved in salt tolerance and methionine biogenesis; dephosphorylates 3'-phosphoadenosine-5'-phosphate and 3'-phosphoadenosine-5'-phosphosulfate, intermediates of the sulfate assimilation pathway |
| MET28 | METhionine requiring | Basic leucine zipper (bZIP) transcriptional activator in the Cbf1p-Met4p-Met28p complex, participates in the regulation of sulfur metabolism |
| MET2 | METhionine requiring | L-homoserine-O-acetyltransferase, catalyzes the conversion of homoserine to O-acetyl homoserine which is the first step of the methionine biosynthetic pathway |
| MET22 | METhionine requiring | Bisphosphate-3'-nucleotidase, involved in salt tolerance and methionine biogenesis; dephosphorylates 3'-phosphoadenosine-5'-phosphate and 3'-phosphoadenosine-5'-phosphosulfate, intermediates of the sulfate assimilation pathway |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| MET28 | METhionine requiring | Basic leucine zipper (bZIP) transcriptional activator in the Cbf1p-Met4p-Met28p complex, participates in the regulation of sulfur metabolism |
| MET30 | METhionine requiring | F-box protein containing five copies of the WD40 motif, controls cell cycle function, sulfur metabolism, and methionine biosynthesis as part of the ubiquitin ligase complex; interacts with and regulates Met4p, localizes within the nucleus |
| MET31 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met32p |
| MET32 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met31p |
| MET3 | METhionine requiring | ATP sulfurylase, catalyzes the primary step of intracellular sulfate activation, essential for assimilatory reduction of sulfate to sulfide, involved in methionine metabolism |
| MET30 | METhionine requiring | F-box protein containing five copies of the WD40 motif, controls cell cycle function, sulfur metabolism, and methionine biosynthesis as part of the ubiquitin ligase complex; interacts with and regulates Met4p, localizes within the nucleus |
| MET31 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met32p |
| MET32 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met31p |
| MET4 | METhionine requiring | Leucine-zipper transcriptional activator, responsible for the regulation of the sulfur amino acid pathway, requires different combinations of the auxiliary factors Cbf1p, Met28p, Met31p and Met32p |
| MOB1 | Mps One Binder | Component of the mitotic exit network; associates with and is required for the activation and Cdc15p-dependent phosphorylation of the Dbf2p kinase; required for cytokinesis and cell separation; component of the CCR4 transcriptional complex |
| MTO1 | Mitochondrial Translation Optimization | Mitochondrial protein, forms a heterodimer complex with Mss1p that performs the 5-carboxymethylaminomethyl modification of the wobble uridine base in mitochondrial tRNAs; required for respiration in paromomycin-resistant 15S rRNA mutants |
| MUP3 | Methionine UPtake | Low affinity methionine permease, similar to Mup1p |
| NAR1 | Nuclear Architecture Related | Component of the cytosolic iron-sulfur (FeS) protein assembly machinery, required for maturation of cytosolic and nuclear FeS proteins and for normal resistance to oxidative stress; homologous to human Narf |
| NBP2 | Nap1 Binding Protein | Protein involved in the HOG (high osmolarity glycerol) pathway, negatively regulates Hog1p by recruitment of phosphatase Ptc1p the Pbs2p-Hog1p complex, found in the nucleus and cytoplasm, contains an SH3 domain that binds Pbs2p |
| NCE103 | NonClassical Export | Carbonic anhydrase; poorly transcribed under aerobic conditions and at an undetectable level under anaerobic conditions; involved in non-classical protein export pathway |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| NIT1 | NITrilase superfamily | Nitrilase, member of the nitrilase branch of the nitrilase superfamily; in closely related species and other S. cerevisiae strain backgrounds YIL164C and adjacent ORF, YIL165C, likely constitute a single ORF encoding a nitrilase gene |
| NIT3 | NITrilase superfamily | Nit protein, one of two proteins in S. cerevisiae with similarity to the Nit domain of NitFhit from fly and worm and to the mouse and human Nit protein which interacts with the Fhit tumor suppressor; nitrilase superfamily member |
| NPR1 | Nitrogen Permease Reactivator | Protein kinase that stabilizes several plasma membrane amino acid transporters by antagonizing their ubiquitin-mediated degradation |
| NRK1 | Nicotinamide Riboside Kinase | Nicotinamide riboside kinase, catalyzes the phosphorylation of nicotinamide riboside and nicotinic acid riboside in salvage pathways for NAD+ biosynthesis |
| NTG1 | eNdonuclease Three-like Glycosylase | DNA N-glycosylase and apurinic/apyrimidinic (AP) lyase involved in base excision repair; distribution between nucleus and mitochondrion varies according to which compartment is under oxidative stress and is also sumoylation-dependent |
| NUP42 | NUclear Pore | Subunit of the nuclear pore complex (NPC) that localizes exclusively to the cytoplasmic side; involved in RNA export, most likely at a terminal step; interacts with Gle1p |
| NYV1 | New Yeast V-SNARE | v-SNARE component of the vacuolar SNARE complex involved in vesicle fusion; inhibits ATP-dependent Ca(2+) transport activity of Pmc1p in the vacuolar membrane |
| ORT1 | ORnithine Transporter | Ornithine transporter of the mitochondrial inner membrane, exports ornithine from mitochondria as part of arginine biosynthesis; human ortholog is associated with hyperammonaemia-hyperornithinaemia-homocitrullinuria (HHH) syndrome |
| PCL10 | Pho85 CycLin | Pho85p cyclin; recruits, activates, and targets Pho85p cyclin-dependent protein kinase to its substrate |
| PCL5 | Pho85 CycLin | Cyclin, interacts with and phosphorylated by Pho85p cyclin-dependent kinase (Cdk), induced by Gcn4p at level of transcription, specifically required for Gcn4p degradation, may be sensor of cellular protein biosynthetic capacity |
| PCL6 | Pho85 CycLin | Pho85p cyclin of the Pho80p subfamily; forms the major Glc8p kinase together with Pcl7p and Pho85p; involved in the control of glycogen storage by Pho85p; stabilized by Elongin C binding |
| PCL8 | Pho85 CycLin | Cyclin, interacts with Pho85p cyclin-dependent kinase (Cdk) to phosphorylate and regulate glycogen synthase, also activates Pho85p for Glc8p phosphorylation |
| PCL9 | Pho85 CycLin | Cyclin, forms a functional kinase complex with Pho85p cyclin-dependent kinase (Cdk), expressed in late M/early G1 phase, activated by Swi5p |
| PDR12 | Pleiotropic Drug Resistance | Plasma membrane ATP-binding cassette (ABC) transporter, weak-acid-inducible multidrug transporter required for weak organic acid resistance; induced by sorbate and benzoate and regulated by War1p; mutants exhibit sorbate hypersensitivity |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| PDX3 | PyriDoXine auxotrophy | Pyridoxine (pyridoxamine) phosphate oxidase, has homologs in E. coli and Myxococcus xanthus; transcription is under the general control of nitrogen metabolism |
| PEX21 | PEroXin | Peroxin required for targeting of peroxisomal matrix proteins containing PTS2; interacts with Pex7p; partially redundant with Pex18p |
| PHO80 | PHOsphate metabolism | Cyclin, interacts with cyclin-dependent kinase Pho85p; regulates the response to nutrient levels and environmental conditions, including the response to phosphate limitation and stress-dependent calcium signaling |
| PHO81 | PHOsphate metabolism | Cyclin-dependent kinase (CDK) inhibitor, regulates Pho80p-Pho85p and Pcl7p-Pho85p cyclin-CDK complexes in response to phosphate levels; inhibitory activity for Pho80p-Pho85p requires myo-D-inositol heptakisphosphate (IP7) generated by Vip1p |
| PHO84 | PHOsphate metabolism | High-affinity inorganic phosphate (Pi) transporter and low-affinity manganese transporter; regulated by Pho4p and Spt7p; mutation confers resistance to arsenate; exit from the ER during maturation requires Pho86p |
| PHO85 | PHOsphate metabolism | Cyclin-dependent kinase, with ten cyclin partners; involved in regulating the cellular response to nutrient levels and environmental conditions and progression through the cell cycle |
| PHO86 | PHOsphate metabolism | Endoplasmic reticulum (ER) resident protein required for ER exit of the high-affinity phosphate transporter Pho84p, specifically required for packaging of Pho84p into COPII vesicles |
| PHO87 | PHOsphate metabolism | Low-affinity inorganic phosphate (Pi) transporter, involved in activation of PHO pathway; expression is independent of Pi concentration and Pho4p activity; contains 12 membrane-spanning segments |
| PHO88 | PHOsphate metabolism | Probable membrane protein, involved in phosphate transport; pho88 pho86 double null mutant exhibits enhanced synthesis of repressible acid phosphatase at high inorganic phosphate concentrations |
| PHO89 | PHOsphate metabolism | Na+/Pi cotransporter, active in early growth phase; similar to phosphate transporters of Neurospora crassa; transcription regulated by inorganic phosphate concentrations and Pho4p |
| PHO8 | PHOsphate metabolism | Repressible alkaline phosphatase, a glycoprotein localized to the vacuole; regulated by levels of inorganic phosphate and by a system consisting of Pho4p, Pho9p, Pho80p, Pho81p and Pho85p; dephosphorylates phosphotyrosyl peptides |
| PHO80 | PHOsphate metabolism | Cyclin, interacts with cyclin-dependent kinase Pho85p; regulates the response to nutrient levels and environmental conditions, including the response to phosphate limitation and stress-dependent calcium signaling |
| PHO81 | PHOsphate metabolism | Cyclin-dependent kinase (CDK) inhibitor, regulates Pho80p-Pho85p and Pcl7p-Pho85p cyclin-CDK complexes in response to phosphate levels; inhibitory activity for Pho80p-Pho85p requires myo-D-inositol heptakisphosphate (IP7) generated by Vip1p |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| PHO84 | PHOsphate metabolism | High-affinity inorganic phosphate (Pi) transporter and low-affinity manganese transporter; regulated by Pho4p and Spt7p; mutation confers resistance to arsenate; exit from the ER during maturation requires Pho86p |
| PHO85 | PHOsphate metabolism | Cyclin-dependent kinase, with ten cyclin partners; involved in regulating the cellular response to nutrient levels and environmental conditions and progression through the cell cycle |
| PHO86 | PHOsphate metabolism | Endoplasmic reticulum (ER) resident protein required for ER exit of the high-affinity phosphate transporter Pho84p, specifically required for packaging of Pho84p into COPII vesicles |
| PHO87 | PHOsphate metabolism | Low-affinity inorganic phosphate (Pi) transporter, involved in activation of PHO pathway; expression is independent of Pi concentration and Pho4p activity; contains 12 membrane-spanning segments |
| PHO88 | PHOsphate metabolism | Probable membrane protein, involved in phosphate transport; pho88 pho86 double null mutant exhibits enhanced synthesis of repressible acid phosphatase at high inorganic phosphate concentrations |
| PHO89 | PHOsphate metabolism | Na+/Pi cotransporter, active in early growth phase; similar to phosphate transporters of Neurospora crassa; transcription regulated by inorganic phosphate concentrations and Pho4p |
| PIP2 | Peroxisome Induction Pathway | Autoregulatory oleate-specific transcriptional activator of peroxisome proliferation, contains Zn(2)-Cys(6) cluster domain, forms heterodimer with Oaf1p, binds oleate response elements (OREs), activates beta-oxidation genes |
| POS5 | PerOxide Sensitive | Mitochondrial NADH kinase, phosphorylates NADH; also phosphorylates NAD(+) with lower specificity; required for the response to oxidative stress |
| PRK1 | p53 Regulatory Kinase | Protein serine/threonine kinase; regulates the organization and function of the actin cytoskeleton through the phosphorylation of the Pan1p-Sla1p-End3p protein complex |
| PUS1 | PseudoUridine Synthase | tRNA:pseudouridine synthase, introduces pseudouridines at positions 26-28, 34-36, 65, and 67 of tRNA; nuclear protein that appears to be involved in tRNA export; also acts on U2 snRNA |
| PUT3 | Proline UTilization | Transcriptional activator of proline utilization genes, constitutively binds PUT1 and PUT2 promoter sequences and undergoes a conformational change to form the active state; has a Zn(2)-Cys(6) binuclear cluster domain |
| RAD55 | RADiation sensitive | Protein that stimulates strand exchange by stabilizing the binding of Rad51p to single-stranded DNA; involved in the recombinational repair of double-strand breaks in DNA during vegetative growth and meiosis; forms heterodimer with Rad57p |
| RAD59 | RADiation sensitive | Protein involved in the repair of double-strand breaks in DNA during vegetative growth via recombination and single-strand annealing; anneals complementary single-stranded DNA; homologous to Rad52p |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| RGT2 | Restores Glucose Transport | Plasma membrane glucose receptor, highly similar to Snf3p; both Rgt2p and Snf3p serve as transmembrane glucose sensors generating an intracellular signal that induces expression of glucose transporter (HXT) genes |
| RIB3 | RIBoflavin biosynthesis | 3,4-dihydroxy-2-butanone-4-phosphate synthase (DHBP synthase), required for riboflavin biosynthesis from ribulose-5-phosphate, also has an unrelated function in mitochondrial respiration |
| RIB5 | RIBoflavin biosynthesis | Riboflavin synthase; catalyzes the last step of the riboflavin biosynthesis pathway |
| RIM101 | Regulator of IME2 | Transcriptional repressor involved in response to pH and in cell wall construction; required for alkaline pH-stimulated haploid invasive growth and sporulation; activated by proteolytic processing; similar to A. nidulans PacC |
| ROT1 | Reversal Of Tor2 lethality | Molecular chaperone involved in protein folding in the ER; mutation causes defects in cell wall synthesis and in lysis of autophagic bodies, suppresses tor2 mutations, and is synthetically lethal with kar2-1 and with rot2 mutations |
| RTG3 | ReTroGrade regulation | Basic helix-loop-helix-leucine zipper (bHLH/Zip) transcription factor that forms a complex with another bHLH/Zip protein, Rtg1p, to activate the retrograde (RTG) and TOR pathways |
| SDL1 | YIL168W Pseudogene | Open reading frame, unlikely to produce a functional protein in S288C; in closely related species and other S. cerevisiae strain backgrounds YIL168W and adjacent ORF, YIL167W, constitute a single ORF encoding L-serine dehydratase |
| SER1 | SERine requiring | 3-phosphoserine aminotransferase, catalyzes the formation of phosphoserine from 3-phosphohydroxypyruvate, required for serine and glycine biosynthesis; regulated by the general control of amino acid biosynthesis mediated by Gcn4p |
| SFT2 | Suppressor of sedFive Ts | Non-essential tetra-spanning membrane protein found mostly in the late Golgi, can suppress some sed5 alleles; may be part of the transport machinery, but precise function is unknown; similar to mammalian syntaxin 5 |
| SIP4 | SNF1-Interacting Protein | C6 zinc cluster transcriptional activator that binds to the carbon source-responsive element (CSRE) of gluconeogenic genes; involved in the positive regulation of gluconeogenesis; regulated by Snf1p protein kinase; localized to the nucleus |
| SIR1 | Silent Information Regulator | Protein involved in repression of transcription at the silent mating-type loci HML and HMR; recruitment to silent chromatin requires interactions with Orc1p and with Sir4p, through a common Sir1p domain; binds to centromeric chromatin |
| SMF3 | | Putative divalent metal ion transporter involved in iron homeostasis; transcriptionally regulated by metal ions; member of the Nramp family of metal transport proteins |
| SNO1 | SNZ proximal Open reading frame | Protein of unconfirmed function, involved in pyridoxine metabolism; expression is induced during stationary phase; forms a putative glutamine amidotransferase complex with Snz1p, with Sno1p serving as the glutaminase |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| SNZ1 | SNooZe | Protein involved in vitamin B6 biosynthesis; member of a stationary phase-induced gene family; coregulated with SNO1; interacts with Sno1p and with Yhr198p, perhaps as a multiprotein complex containing other Snz and Sno proteins |
| SPO19 | SPOrulation | Meiosis-specific prospore protein; required to produce bending force necessary for proper assembly of the prospore membrane during sporulation; identified as a weak high-copy suppressor of the spo1-1 ts mutation |
| SRB8 | Suppressor of RNA polymerase B | Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation; involved in glucose repression |
| STB4 | Sin Three Binding protein | Protein that binds Sin3p in a two-hybrid assay; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters |
| STP1 | Species-specific tRNA Processing | Transcription factor, undergoes proteolytic processing by SPS (Ssy1p-Ptr3p-Ssy5p)-sensor component Ssy5p in response to extracellular amino acids; activates transcription of amino acid permease genes and may have a role in tRNA processing |
| STP22 | STerile Pseudoreversion | Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype |
| STP2 | protein with similarity to Stp1p | Transcription factor, activated by proteolytic processing in response to signals from the SPS sensor system for external amino acids; activates transcription of amino acid permease genes |
| STP22 | STerile Pseudoreversion | Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype |
| SUL2 | SULfate metabolism | High affinity sulfate permease; sulfate uptake is mediated by specific sulfate transporters Sul1p and Sul2p, which control the concentration of endogenous activated sulfate intermediates |
| SUT1 | Sterol UpTake | Transcription factor of the Zn[II]2Cys6 family involved in sterol uptake; involved in induction of hypoxic gene expression |
| TAH18 | Top1T722A mutant Hypersensitive | Protein of unknown function that plays a pro-death role in response to oxidative stress; highly conserved across species and similar to human protein NDOR1; allele is synthetically lethal with the pol3-13 allele of DNA polymerase delta |
| TEA1 | Ty Enhancer Activator | Ty1 enhancer activator required for full levels of Ty enhancer-mediated transcription; C6 zinc cluster DNA-binding protein |
| THR4 | THReonine requiring | Threonine synthase, conserved protein that catalyzes formation of threonine from 0-phosphohomoserine; expression is regulated by the GCN4-mediated general amino acid control pathway |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| TRP2 | TRyPtophan requiring | Anthranilate synthase, catalyzes the initial step of tryptophan biosynthesis, forms multifunctional hetero-oligomeric anthranilate synthase:indole-3-glycerol phosphate synthase enzyme complex with Trp3p |
| TRP3 | TRyPtophan requiring | Bifunctional enzyme exhibiting both indole-3-glycerol-phosphate synthase and anthranilate synthase activities, forms multifunctional hetero-oligomeric anthranilate synthase:indole-3-glycerol phosphate synthase enzyme complex with Trp2p |
| TRP4 | TRyPtophan requiring | Anthranilate phosphoribosyl transferase of the tryptophan biosynthetic pathway, catalyzes the phosphoribosylation of anthranilate, subject to the general control system of amino acid biosynthesis |
| TRP5 | TRyPtophan requiring | Tryptophan synthase, catalyzes the last step of tryptophan biosynthesis; regulated by the general control system of amino acid biosynthesis |
| UGA3 | Utilization of GAba | Transcriptional activator necessary for gamma-aminobutyrate (GABA)-dependent induction of GABA genes (such as UGA1, UGA2, UGA4); zinc-finger transcription factor of the Zn(2)-Cys(6) binuclear cluster domain type; localized to the nucleus |
| UGX2 | Unidentified Gene X | Protein of unknown function, transcript accumulates in response to any combination of stress conditions |
| URA10 | URAcil requiring | Minor orotate phosphoribosyltransferase (OPRTase) isozyme that catalyzes the fifth enzymatic step in the de novo biosynthesis of pyrimidines, converting orotate into orotidine-5'-phosphate; major OPRTase encoded by URA5 |
| URE2 | UREidosuccinate transport | Nitrogen catabolite repression transcriptional regulator that acts by inhibition of GLN3 transcription in good nitrogen source; has glutathione peroxidase activity and can mutate to acquire GST activity; altered form creates [URE3] prion |
| VPS27 | Vacuolar Protein Sorting | Endosomal protein that forms a complex with Hse1p; required for recycling Golgi proteins, forming lumenal membranes and sorting ubiquitinated proteins destined for degradation; has Ubiquitin Interaction Motifs which bind ubiquitin (Ubi4p) |
| VPS34 | Vacuolar Protein Sorting | Phosphatidylinositol 3-kinase responsible for the synthesis of phosphatidylinositol 3-phosphate; forms membrane-associated signal transduction complex with Vps15p to regulate protein sorting; activated by the GTP-bound form of Gpa1p |
| VPS41 | Vacuolar Protein Sorting | Vacuolar membrane protein that is a subunit of the homotypic vacuole fusion and vacuole protein sorting (HOPS) complex; essential for membrane docking and fusion at the Golgi-to-endosome and endosome-to-vacuole stages of protein transport |
| XDJ1 | Putative chaperone | Putative chaperone, homolog of E. coli DnaJ, closely related to Ydj1p; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies |
| YBT1 | Yeast Bile Transporter | Transporter of the ATP-binding cassette (ABC) family involved in bile acid transport; similar to mammalian bile transporters |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| YCK3 | Yeast Casein Kinase | Palmitoylated, vacuolar membrane-localized casein kinase I isoform; negatively regulates vacuole fusion during hypertonic stress via phosphorylation of Vps41p; shares essential functions with Hrr25p; regulates vesicle fusion in AP-3 pathway |
| YCR023C | N/A | Vacuolar membrane protein of unknown function; member of the multidrug resistance family; YCR023C is not an essential gene |
| YCR100C | N/A | Putative protein of unknown function |
| YDL025C | N/A | Putative protein kinase, potentially phosphorylated by Cdc28p; YDL025C is not an essential gene |
| YDL057W | N/A | Putative protein of unknown function; YDL057W is not an essential gene |
| YDR090C | N/A | Putative protein of unknown function |
| YDR186C | N/A | Putative protein of unknown function; may interact with ribosomes, based on co-purification experiments; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm |
| YDR249C | N/A | Putative protein of unknown function |
| YDR306C | N/A | F-box protein of unknown function; interacts with Sgt1p via a Leucine-Rich Repeat (LRR) domain |
| YDR531W | Coenzyme A Biosynthesis | Pantothenate kinase (ATP:D-pantothenate 4'-phosphotransferase, EC 2.7.1.33); catalyzes the first committed step in the universal biosynthetic pathway for synthesis of coenzyme A (CoA) |
| YEL045C | N/A | Dubious open reading frame unlikely to encode a protein based on available experimental and comparative sequence data; deletion gives MMS sensitivity, growth defect under alkaline conditions, less than optimal growth upon citric acid stress |
| YER128W | N/A | Putative protein of unknown function |
| YGL059W | Protein Kinase of PDH | Mitochondrial protein kinase that negatively regulates activity of the pyruvate dehydrogenase complex by phosphorylating the ser-133 residue of the Pda1p subunit; acts in concert with kinase Pkp1p and phosphatases Ptc5p and Ptc6p |
| YGL114W | N/A | Putative protein of unknown function; predicted member of the oligopeptide transporter (OPT) family of membrane transporters |
| YGL117W | N/A | Putative protein of unknown function |
| YHR112C | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm |
| YHR122W | N/A | Protein of unknown function required for establishment of sister chromatid cohesion; synthetically lethal with RFC5, an RF-C subunit that links replication to cohesion establishment; YHR122W is an essential gene |
| YHR162W | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the mitochondrion |
| YIL166C | N/A | Putative protein with similarity to the allantoate permease (Dal5p) subfamily of the major facilitator superfamily; mRNA expression is elevated by sulfur limitation; YIL166C is a non-essential gene |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| YJL206C | N/A | Putative protein of unknown function; similar to transcriptional regulators from the Zn[2]-Cys[6] binuclear cluster protein family; mRNA is weakly cell cycle regulated, peaking in S phase; induced rapidly upon MMS treatment |
| YJR111C | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the mitochondria |
| YJR149W | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm |
| YKL023W | N/A | Putative protein of unknown function, predicted by computational methods to be involved in mRNA degradation; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm |
| YKL107W | N/A | Putative protein of unknown function; proposed to be a palmitoylated membrane protein |
| YKR033C | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene DAL80 |
| YKR096W | Est1A/B-Smg6/5-Like | Protein of unknown function that may interact with ribosomes, based on co-purification experiments; green fluorescent protein (GFP)-fusion protein localizes to the nucleus and cytoplasm; predicted to contain a PINc domain |
| YLL029W | Fe Repressor of Activation | Protein involved in negative regulation of transcription of iron regulon; forms an iron independent complex with Fra2p, Grx3p, and Grx4p; cytosolic; mutant fails to repress transcription of iron regulon and is defective in spore formation |
| YLR152C | N/A | Putative protein of unknown function; YLR152C is not an essential gene |
| YLR302C | N/A | Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data |
| YMC1 | Yeast Mitochondrial Carrier | Mitochondrial protein, putative inner membrane transporter with a role in oleate metabolism and glutamate biosynthesis; member of the mitochondrial carrier (MCF) family; has similarity with Ymc2p |
| YMC2 | Yeast Mitochondrial Carrier | Mitochondrial protein, putative inner membrane transporter with a role in oleate metabolism and glutamate biosynthesis; member of the mitochondrial carrier (MCF) family; has similarity with Ymc1p |
| YMR321C | N/A | Putative protein of unknown function; proposed to be a palmitoylated membrane protein |
| YNL276C | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; overlaps the verified gene MET2/YNL277W |
| YNR066C | N/A | Putative membrane-localized protein of unknown function |
| YOL118C | N/A | Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data |
| YOR203W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; overlaps 5' end of essential DED1 gene required for translation initiation |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 6** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| YOR225W | N/A | Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data |
| YPL033C | Suppressor of Rad53 null Lethality or Suppressor of Rad53 and Lcd1 | Protein of unknown function; involved in regulation of dNTP production; null mutant suppresses the lethality of lcd1 and rad53 mutations; expression is induced by Kar4p |
| YPL264C | N/A | Putative membrane protein of unknown function; physically interacts with Hsp82p; YPL264C is not an essential gene |
| YPR059C | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene YMC1/YPR058W |
| YPR078C | N/A | Putative protein of unknown function; possible role in DNA metabolism and/or in genome stability; expression is heat-inducible |
| YSA1 | N/A | Nudix hydrolase family member with ADP-ribose pyrophosphatase activity; shown to metabolize O-acetyl-ADP-ribose to AMP and acetylated ribose 5'-phosphate |
| YSC83 | N/A | Non-essential mitochondrial protein of unknown function; mRNA induced during meiosis, peaking between mid to late prophase of meiosis I; similar to S. douglasii YSD83 |
| ZTA1 | ZeTA-crystallin | NADPH-dependent quinone reductase, GFP-tagged protein localizes to the cytoplasm and nucleus; has similarity to E. coli quinone oxidoreductase and to human zeta-crystallin |

Table A.4: **Genes in putative amino-acid biosynthesis and -metabolism cluster 9 formed by JFuzzy-K from data of** Brem *et al.* **(2002).**

| Name | Short description | Full description |
|---|---|---|
| AAD10 | Aryl-Alcohol Dehydrogenase | Putative aryl-alcohol dehydrogenase with similarity to P. chrysosporium aryl-alcohol dehydrogenase; mutational analysis has not yet revealed a physiological role |
| AAT1 | Aspartate AminoTransferase | Mitochondrial aspartate aminotransferase, catalyzes the conversion of oxaloacetate to aspartate in aspartate and asparagine biosynthesis |
| AAT2 | Aspartate AminoTransferase | Cytosolic aspartate aminotransferase, involved in nitrogen metabolism; localizes to peroxisomes in oleate-grown cells |
| ADE3 | ADEnine requiring | Cytoplasmic trifunctional enzyme C1-tetrahydrofolate synthase, involved in single carbon metabolism and required for biosynthesis of purines, thymidylate, methionine, and histidine; null mutation causes auxotrophy for adenine and histidine |
| ADE4 | ADEnine requiring | Phosphoribosylpyrophosphate amidotransferase (PRPPAT; amidophosphoribosyltransferase), catalyzes first step of the 'de novo' purine nucleotide biosynthetic pathway |
| ADH5 | Alcohol DeHydrogenase | Alcohol dehydrogenase isoenzyme V; involved in ethanol production |
| ALD5 | ALdehyde Dehydrogenase | Mitochondrial aldehyde dehydrogenase, involved in regulation or biosynthesis of electron transport chain components and acetate formation; activated by K+; utilizes NADP+ as the preferred coenzyme; constitutively expressed |
| ARO1 | AROmatic amino acid requiring | Pentafunctional arom protein, catalyzes steps 2 through 6 in the biosynthesis of chorismate, which is a precursor to aromatic amino acids |
| ARO2 | AROmatic amino acid requiring | Bifunctional chorismate synthase and flavin reductase, catalyzes the conversion of 5-enolpyruvylshikimate 3-phosphate (EPSP) to form chorismate, which is a precursor to aromatic amino acids |
| ARO4 | AROmatic amino acid requiring | 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) synthase, catalyzes the first step in aromatic amino acid biosynthesis and is feedback-inhibited by tyrosine or high concentrations of phenylalanine or tryptophan |
| ASN1 | ASparagiNe requiring | Asparagine synthetase, isozyme of Asn2p; catalyzes the synthesis of L-asparagine from L-aspartate in the asparagine biosynthetic pathway |
| ATR1 | AminoTriazole Resistance | Multidrug efflux pump of the major facilitator superfamily, required for resistance to aminotriazole and 4-nitroquinoline-N-oxide |
| BAP2 | Branched-chain Amino acid Permease | High-affinity leucine permease, functions as a branched-chain amino acid permease involved in the uptake of leucine, isoleucine and valine; contains 12 predicted transmembrane domains |
| BIG1 | Bad In Glucose | Integral membrane protein of the endoplasmic reticulum, required for normal content of cell wall beta-1,6-glucan |
| CAN1 | CANavanine resistance | Plasma membrane arginine permease, requires phosphatidyl ethanolamine (PE) for localization, exclusively associated with lipid rafts; mutation confers canavanine resistance |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| CIT2 | CITrate synthase | Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle; expression is controlled by Rtg1p and Rtg2p transcription factors |
| CLG1 | Cyclin-Like Gene | Cyclin-like protein that interacts with Pho85p; has sequence similarity to G1 cyclins PCL1 and PCL2 |
| COT1 | CObalt Toxicity | Vacuolar transporter that mediates zinc transport into the vacuole; overexpression confers resistance to cobalt and rhodium |
| CRZ1 | Calcineurin-Responsive Zinc finger | Transcription factor that activates transcription of genes involved in stress response; nuclear localization is positively regulated by calcineurin-mediated dephosphorylation |
| CTF13 | Chromosome Transmission Fidelity | Subunit of the CBF3 complex, which binds to the CDE III element of centromeres, bending the DNA upon binding, and may be involved in sister chromatid cohesion during mitosis |
| DIC1 | DIcarboxylate Carrier | Mitochondrial dicarboxylate carrier, integral membrane protein, catalyzes a dicarboxylate-phosphate exchange across the inner mitochondrial membrane, transports cytoplasmic dicarboxylates into the mitochondrial matrix |
| DST1 | DNA Strand Transfer | General transcription elongation factor TFIIS, enables RNA polymerase II to read through blocks to elongation by stimulating cleavage of nascent transcripts stalled at transcription arrest sites |
| ECM17 | METhionine requiring | Sulfite reductase beta subunit, involved in amino acid biosynthesis, transcription repressed by methionine |
| ERG8 | ERGosterol biosynthesis | Phosphomevalonate kinase, an essential cytosolic enzyme that acts in the biosynthesis of isoprenoids and sterols, including ergosterol, from mevalonate |
| ERV1 | Essential for Respiration and Viability | Flavin-linked sulfhydryl oxidase of the mitochondrial intermembrane space (IMS), oxidizes Mia40p as part of a disulfide relay system that promotes IMS retention of imported proteins; ortholog of human hepatopoietin (ALR) |
| ERV14 | ER Vesicle | Protein localized to COPII-coated vesicles, involved in vesicle formation and incorporation of specific secretory cargo; required for the delivery of bud-site selection protein Axl2p to cell surface; related to Drosophila cornichon |
| ESBP6 | MCH3 | Protein with similarity to monocarboxylate permeases, appears not to be involved in transport of monocarboxylates such as lactate, pyruvate or acetate across the plasma membrane |
| FOL2 | FOLic acid synthesis | GTP-cyclohydrolase I, catalyzes the first step in the folic acid biosynthetic pathway |
| FRE6 | Ferric REductase | Putative ferric reductase with similarity to Fre2p; expression induced by low iron levels |
| GDH2 | Glutamate DeHydrogenase | NAD(+)-dependent glutamate dehydrogenase, degrades glutamate to ammonia and alpha-ketoglutarate; expression sensitive to nitrogen catabolite repression and intracellular ammonia levels |
| GLY1 | GLYcine requiring | Threonine aldolase, catalyzes the cleavage of L-allo-threonine and L-threonine to glycine; involved in glycine biosynthesis |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| GPD2 | Glycerol-3-Phosphate Dehydrogenase | NAD-dependent glycerol 3-phosphate dehydrogenase, homolog of Gpd1p, expression is controlled by an oxygen-independent signaling pathway required to regulate metabolism under anoxic conditions; located in cytosol and mitochondria |
| GTR2 | GTp binding protein Resemblance | Putative GTP binding protein that negatively regulates Ran/Tc4 GTPase cycle; activates transcription; subunit of EGO and GSE complexes; required for sorting of Gap1p; localizes to cytoplasm and to chromatin; homolog of human RagC and RagD |
| HIS1 | HIStidine requiring | ATP phosphoribosyltransferase, a hexameric enzyme, catalyzes the first step in histidine biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts; transcription is regulated by general amino acid control |
| HIS3 | HIStidine requiring | Imidazoleglycerol-phosphate dehydratase, catalyzes the sixth step in histidine biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts; transcription is regulated by general amino acid control via Gcn4p |
| HIS5 | HIStidine requiring | Histidinol-phosphate aminotransferase, catalyzes the seventh step in histidine biosynthesis; responsive to general control of amino acid biosynthesis; mutations cause histidine auxotrophy and sensitivity to Cu, Co, and Ni salts |
| HIS7 | HIStidine requiring | Imidazole glycerol phosphate synthase (glutamine amidotransferase:cyclase), catalyzes the fifth and sixth steps of histidine biosynthesis and also produces 5-aminoimidazole-4-carboxamide ribotide (AICAR), a purine precursor |
| HOM2 | HOMoserine requiring | Aspartic beta semi-aldehyde dehydrogenase, catalyzes the second step in the common pathway for methionine and threonine biosynthesis; expression regulated by Gcn4p and the general control of amino acid synthesis |
| HOM3 | HOMoserine requiring | Aspartate kinase (L-aspartate 4-P-transferase); cytoplasmic enzyme that catalyzes the first step in the common pathway for methionine and threonine biosynthesis; expression regulated by Gcn4p and the general control of amino acid synthesis |
| HRB1 | Hypothetical RNA-Binding protein | Poly(A+) RNA-binding protein, involved in the export of mRNAs from the nucleus to the cytoplasm; similar to Gbp2p and Npl3p |
| IDP1 | Isocitrate Dehydrogenase, NADP-specific | Mitochondrial NADP-specific isocitrate dehydrogenase, catalyzes the oxidation of isocitrate to alpha-ketoglutarate; not required for mitochondrial respiration and may function to divert alpha-ketoglutarate to biosynthetic processes |
| ILV3 | IsoLeucine-plus-Valine requiring | Dihydroxyacid dehydratase, catalyzes third step in the common pathway leading to biosynthesis of branched-chain amino acids |
| ILV5 | IsoLeucine-plus-Valine requiring | Acetohydroxyacid reductoisomerase, mitochondrial protein involved in branched-chain amino acid biosynthesis, also required for maintenance of wild-type mitochondrial DNA and found in mitochondrial nucleoids |
| ILV6 | IsoLeucine-plus-Valine requiring | Regulatory subunit of acetolactate synthase, which catalyzes the first step of branched-chain amino acid biosynthesis; enhances activity of the Ilv2p catalytic subunit, localizes to mitochondria |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| ISU2 | Iron-Sulfur cluster nifU-like protein | Conserved protein of the mitochondrial matrix, required for synthesis of mitochondrial and cytosolic iron-sulfur proteins, performs a scaffolding function in mitochondria during Fe/S cluster assembly; isu1 isu2 double mutant is inviable |
| LEU4 | LEUcine biosynthesis | Alpha-isopropylmalate synthase (2-isopropylmalate synthase); the main isozyme responsible for the first step in the leucine biosynthesis pathway |
| LYS4 | LYSine requiring | Homoaconitase, catalyzes the conversion of homocitrate to homoisocitrate, which is a step in the lysine biosynthesis pathway |
| MAE1 | MAlic Enzyme | Mitochondrial malic enzyme, catalyzes the oxidative decarboxylation of malate to pyruvate, which is a key intermediate in sugar metabolism and a precursor for synthesis of several amino acids |
| MCT1 | Malonyl-CoA:ACP Transferase | Predicted malonyl-CoA:ACP transferase, putative component of a type-II mitochondrial fatty acid synthase that produces intermediates for phospholipid remodeling |
| MET1 | METhionine requiring | S-adenosyl-L-methionine uroporphyrinogen III transmethylase, involved in the biosynthesis of siroheme, a prosthetic group used by sulfite reductase; required for sulfate assimilation and methionine biosynthesis |
| MET10 | METhionine requiring | Subunit alpha of assimilatory sulfite reductase, which converts sulfite into sulfide |
| MET12 | METhionine requiring | Protein with methylenetetrahydrofolate reductase (MTHFR) activity in vitro; null mutant has no phenotype and is prototrophic for methionine; MET13 encodes major isozyme of MTHFR |
| MET13 | METhionine requiring | Major isozyme of methylenetetrahydrofolate reductase, catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate in the methionine biosynthesis pathway |
| MET14 | METhionine requiring | Adenylylsulfate kinase, required for sulfate assimilation and involved in methionine metabolism |
| MET16 | METhionine requiring | 3'-phosphoadenylsulfate reductase, reduces 3'-phosphoadenylyl sulfate to adenosine-3',5'-bisphosphate and free sulfite using reduced thioredoxin as cosubstrate, involved in sulfate assimilation and methionine metabolism |
| MET17 | METhionine requiring | Methionine and cysteine synthase (O-acetyl homoserine-O-acetyl serine sulfhydrylase), required for sulfur amino acid synthesis |
| MET18 | METhionine requiring | DNA repair and TFIIH regulator, required for both nucleotide excision repair (NER) and RNA polymerase II (RNAP II) transcription; involved in telomere maintenance |
| MET2 | METhionine requiring | L-homoserine-O-acetyltransferase, catalyzes the conversion of homoserine to O-acetyl homoserine which is the first step of the methionine biosynthetic pathway |
| MET22 | METhionine requiring | Bisphosphate-3'-nucleotidase, involved in salt tolerance and methionine biogenesis; dephosphorylates 3'-phosphoadenosine-5'-phosphate and 3'-phosphoadenosine-5'-phosphosulfate, intermediates of the sulfate assimilation pathway |
| MET28 | METhionine requiring | Basic leucine zipper (bZIP) transcriptional activator in the Cbf1p-Met4p-Met28p complex, participates in the regulation of sulfur metabolism |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| MET3 | METhionine requiring | ATP sulfurylase, catalyzes the primary step of intracellular sulfate activation, essential for assimilatory reduction of sulfate to sulfide, involved in methionine metabolism |
| MET30 | METhionine requiring | F-box protein containing five copies of the WD40 motif, controls cell cycle function, sulfur metabolism, and methionine biosynthesis as part of the ubiquitin ligase complex; interacts with and regulates Met4p, localizes within the nucleus |
| MET31 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met32p |
| MET32 | METhionine requiring | Zinc-finger DNA-binding protein, involved in transcriptional regulation of the methionine biosynthetic genes, similar to Met31p |
| NAR1 | Nuclear Architecture Related | Component of the cytosolic iron-sulfur (FeS) protein assembly machinery, required for maturation of cytosolic and nuclear FeS proteins and for normal resistance to oxidative stress; homologous to human Narf |
| NCE103 | NonClassical Export | Carbonic anhydrase; poorly transcribed under aerobic conditions and at an undetectable level under anaerobic conditions; involved in non-classical protein export pathway |
| NFS1 | NiFS-like | Cysteine desulfurase involved in iron-sulfur cluster (Fe/S) biogenesis; required for the post-transcriptional thio-modification of mitochondrial and cytoplasmic tRNAs; essential protein located predominantly in mitochondria |
| PCL10 | Pho85 CycLin | Pho85p cyclin; recruits, activates, and targets Pho85p cyclin-dependent protein kinase to its substrate |
| PCL5 | Pho85 CycLin | Cyclin, interacts with and phosphorylated by Pho85p cyclin-dependent kinase (Cdk), induced by Gcn4p at level of transcription, specifically required for Gcn4p degradation, may be sensor of cellular protein biosynthetic capacity |
| PCL6 | Pho85 CycLin | Pho85p cyclin of the Pho80p subfamily; forms the major Glc8p kinase together with Pcl7p and Pho85p; involved in the control of glycogen storage by Pho85p; stabilized by Elongin C binding |
| PCL7 | Pho85 CycLin | Pho85p cyclin of the Pho80p subfamily, forms a functional kinase complex with Pho85p which phosphorylates Mmr1p and is regulated by Pho81p; involved in glycogen metabolism, expression is cell-cycle regulated |
| PCL8 | Pho85 CycLin | Cyclin, interacts with Pho85p cyclin-dependent kinase (Cdk) to phosphorylate and regulate glycogen synthase, also activates Pho85p for Glc8p phosphorylation |
| PCL9 | Pho85 CycLin | Cyclin, forms a functional kinase complex with Pho85p cyclin-dependent kinase (Cdk), expressed in late M/early G1 phase, activated by Swi5p |
| PDR12 | Pleiotropic Drug Resistance | Plasma membrane ATP-binding cassette (ABC) transporter, weak-acid-inducible multidrug transporter required for weak organic acid resistance; induced by sorbate and benzoate and regulated by War1p; mutants exhibit sorbate hypersensitivity |
| PDX3 | PyriDoXine auxotrophy | Pyridoxine (pyridoxamine) phosphate oxidase, has homologs in E. coli and Myxococcus xanthus; transcription is under the general control of nitrogen metabolism |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| PHO8 | PHOsphate metabolism | Repressible alkaline phosphatase, a glycoprotein localized to the vacuole; regulated by levels of inorganic phosphate and by a system consisting of Pho4p, Pho9p, Pho80p, Pho81p and Pho85p; dephosphorylates phosphotyrosyl peptides |
| PHO80 | PHOsphate metabolism | Cyclin, interacts with cyclin-dependent kinase Pho85p; regulates the response to nutrient levels and environmental conditions, including the response to phosphate limitation and stress-dependent calcium signaling |
| PHO81 | PHOsphate metabolism | Cyclin-dependent kinase (CDK) inhibitor, regulates Pho80p-Pho85p and Pcl7p-Pho85p cyclin-CDK complexes in response to phosphate levels; inhibitory activity for Pho80p-Pho85p requires myo-D-inositol heptakisphosphate (IP7) generated by Vip1p |
| PHO84 | PHOsphate metabolism | High-affinity inorganic phosphate (Pi) transporter and low-affinity manganese transporter; regulated by Pho4p and Spt7p; mutation confers resistance to arsenate; exit from the ER during maturation requires Pho86p |
| PHO85 | PHOsphate metabolism | Cyclin-dependent kinase, with ten cyclin partners; involved in regulating the cellular response to nutrient levels and environmental conditions and progression through the cell cycle |
| PHO86 | PHOsphate metabolism | Endoplasmic reticulum (ER) resident protein required for ER exit of the high-affinity phosphate transporter Pho84p, specifically required for packaging of Pho84p into COPII vesicles |
| PHO87 | PHOsphate metabolism | Low-affinity inorganic phosphate (Pi) transporter, involved in activation of PHO pathway; expression is independent of Pi concentration and Pho4p activity; contains 12 membrane-spanning segments |
| PHO88 | PHOsphate metabolism | Probable membrane protein, involved in phosphate transport; pho88 pho86 double null mutant exhibits enhanced synthesis of repressible acid phosphatase at high inorganic phosphate concentrations |
| PHO89 | PHOsphate metabolism | Na+/Pi cotransporter, active in early growth phase; similar to phosphate transporters of Neurospora crassa; transcription regulated by inorganic phosphate concentrations and Pho4p |
| POS5 | PerOxide Sensitive | Mitochondrial NADH kinase, phosphorylates NADH; also phosphorylates NAD(+) with lower specificity; required for the response to oxidative stress |
| RGA1 | Rho GTPase Activating Protein | GTPase-activating protein for the polarity-establishment protein Cdc42p; implicated in control of septin organization, pheromone response, and haploid invasive growth |
| RGT2 | Restores Glucose Transport | Plasma membrane glucose receptor, highly similar to Snf3p; both Rgt2p and Snf3p serve as transmembrane glucose sensors generating an intracellular signal that induces expression of glucose transporter (HXT) genes |
| RIB3 | RIBoflavin biosynthesis | 3,4-dihydroxy-2-butanone-4-phosphate synthase (DHBP synthase), required for riboflavin biosynthesis from ribulose-5-phosphate, also has an unrelated function in mitochondrial respiration |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| RTG2 | ReTroGrade regulation | Sensor of mitochondrial dysfunction; regulates the subcellular location of Rtg1p and Rtg3p, transcriptional activators of the retrograde (RTG) and TOR pathways; Rtg2p is inhibited by the phosphorylated form of Mks1p |
| RTG3 | ReTroGrade regulation | Basic helix-loop-helix-leucine zipper (bHLH/Zip) transcription factor that forms a complex with another bHLH/Zip protein, Rtg1p, to activate the retrograde (RTG) and TOR pathways |
| SDL1 | YIL168W Pseudogene | Open reading frame, unlikely to produce a functional protein in S288C; in closely related species and other S. cerevisiae strain backgrounds YIL168W and adjacent ORF, YIL167W, constitute a single ORF encoding L-serine dehydratase |
| SMF3 | N/A | Putative divalent metal ion transporter involved in iron homeostasis; transcriptionally regulated by metal ions; member of the Nramp family of metal transport proteins |
| SNO1 | SNZ proximal Open reading frame | Protein of unconfirmed function, involved in pyridoxine metabolism; expression is induced during stationary phase; forms a putative glutamine amidotransferase complex with Snz1p, with Sno1p serving as the glutaminase |
| SNZ1 | SNooZe | Protein involved in vitamin B6 biosynthesis; member of a stationary phase-induced gene family; coregulated with SNO1; interacts with Sno1p and with Yhr198p, perhaps as a multiprotein complex containing other Snz and Sno proteins |
| SOL1 | Suppressor Of Los1-1 | Protein with a possible role in tRNA export; shows similarity to 6-phosphogluconolactonase non-catalytic domains but does not exhibit this enzymatic activity; homologous to Sol2p, Sol3p, and Sol4p |
| SPE1 | SPErmidine auxotroph | Ornithine decarboxylase, catalyzes the first step in polyamine biosynthesis; degraded in a proteasome-dependent manner in the presence of excess polyamines |
| SPO19 | SPOrulation | Meiosis-specific prospore protein; required to produce bending force necessary for proper assembly of the prospore membrane during sporulation; identified as a weak high-copy suppressor of the spo1-1 ts mutation |
| STB4 | Sin Three Binding protein | Protein that binds Sin3p in a two-hybrid assay; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters |
| STP2 | protein with similarity to Stp1p | Transcription factor, activated by proteolytic processing in response to signals from the SPS sensor system for external amino acids; activates transcription of amino acid permease genes |
| STP22 | STerile Pseudoreversion | Component of the ESCRT-I complex, which is involved in ubiquitin-dependent sorting of proteins into the endosome; homologous to the mouse and human Tsg101 tumor susceptibility gene; mutants exhibit a Class E Vps phenotype |
| SUL2 | SULfate metabolism | High affinity sulfate permease; sulfate uptake is mediated by specific sulfate transporters Sul1p and Sul2p, which control the concentration of endogenous activated sulfate intermediates |

*continued on the next page*

**Genes in putative amino-acid biosynthesis and -metabolism cluster 9** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| TAH11 | Topo-A Hypersensitive | DNA replication licensing factor, required for pre-replication complex assembly |
| TAH18 | Top1T722A mutant Hypersensitive | Protein of unknown function that plays a pro-death role in response to oxidative stress; highly conserved across species and similar to human protein NDOR1; allele is synthetically lethal with the pol3-13 allele of DNA polymerase delta |
| THR1 | THReonine requiring | Homoserine kinase, conserved protein required for threonine biosynthesis; expression is regulated by the GCN4-mediated general amino acid control pathway |
| TRP3 | TRyPtophan requiring | Bifunctional enzyme exhibiting both indole-3-glycerol-phosphate synthase and anthranilate synthase activities, forms multifunctional hetero-oligomeric anthranilate synthase:indole-3-glycerol phosphate synthase enzyme complex with Trp2p |
| TRP4 | TRyPtophan requiring | Anthranilate phosphoribosyl transferase of the tryptophan biosynthetic pathway, catalyzes the phosphoribosylation of anthranilate, subject to the general control system of amino acid biosynthesis |
| TRP5 | TRyPtophan requiring | Tryptophan synthase, catalyzes the last step of tryptophan biosynthesis; regulated by the general control system of amino acid biosynthesis |
| URE2 | UREidosuccinate transport | Nitrogen catabolite repression transcriptional regulator that acts by inhibition of GLN3 transcription in good nitrogen source; has glutathione peroxidase activity and can mutate to acquire GST activity; altered form creates [URE3] prion |
| XDJ1 | Putative chaperone | Putative chaperone, homolog of E. coli DnaJ, closely related to Ydj1p; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies |
| YBT1 | Yeast Bile Transporter | Transporter of the ATP-binding cassette (ABC) family involved in bile acid transport; similar to mammalian bile transporters |
| YMC2 | Yeast Mitochondrial Carrier | Mitochondrial protein, putative inner membrane transporter with a role in oleate metabolism and glutamate biosynthesis; member of the mitochondrial carrier (MCF) family; has similarity with Ymc1p |

Table A.5: **Genes in putative mating cluster 1 formed by JFuzzy-K from data of Brem *et al.* (2002).**

| Name | Short description | Full description |
|---|---|---|
| AKR2 | AnKyrin Repeat-containing protein | Ankyrin repeat-containing protein similar to Akr1p; member of a family of putative palmitoyltransferases containing an Asp-His-His-Cys-cysteine rich (DHHC-CRD) domain; possibly involved in constitutive endocytosis of Ste3p |
| HML$\alpha$1 | Hidden Mat Left $\alpha$ | Silenced copy of $\alpha_1$ at HML, encoding a transcriptional coactivator involved in the regulation of mating-type alpha-specific gene expression |
| HML$\alpha$2 | Hidden Mat Left $\alpha$ | Silenced copy of $\alpha_2$ at HML; homeobox-domain protein that associates with Mcm1p in haploid cells to repress a-specific gene expression and interacts with a1p in diploid cells to repress haploid-specific gene expression |
| LEU2 | LEUcine biosynthesis | Beta-isopropylmalate dehydrogenase (IMDH), catalyzes the third step in the leucine biosynthesis pathway |
| LIF1 | Ligase Interacting Factor | Component of the DNA ligase IV complex that mediates nonhomologous end joining in DNA double-strand break repair; physically interacts with Dnl4p and Nej1p; homologous to mammalian XRCC4 protein |
| MAT$\alpha$1 | MATing type protein $\alpha$ | Transcriptional co-activator involved in regulation of mating-type-specific gene expression; targets the transcription factor Mcm1p to the promoters of alpha-specific genes; one of two genes encoded by the MATalpha mating type cassette |
| MAT$\alpha$2 | MATing type protein $\alpha$ | Homeobox-domain protein that, with Mcm1p, represses a-specific genes in haploids; acts with A1p to repress transcription of haploid-specific genes in diploids; one of two genes encoded by the MATalpha mating type cassette |
| MF($\alpha$)1 | Mating Factor $\alpha$ | Mating pheromone alpha-factor, made by alpha cells; interacts with mating type a cells to induce cell cycle arrest and other responses leading to mating; also encoded by MF(ALPHA)2, although MF(ALPHA)1 produces most alpha-factor |
| MF($\alpha$)2 | Mating Factor $\alpha$ | Mating pheromone alpha-factor, made by alpha cells; interacts with mating type a cells to induce cell cycle arrest and other responses leading to mating; also encoded by MF(ALPHA)1, which is more highly expressed than MF(ALPHA)2 |
| SAG1 | Sexual AGglutination | Alpha-agglutinin of alpha-cells, binds to Aga1p during agglutination, N-terminal half is homologous to the immunoglobulin superfamily and contains binding site for a-agglutinin, C-terminal half is highly glycosylated and contains GPI anchor |
| STE3 | STErile | Receptor for a factor pheromone, transcribed in alpha cells and required for mating by alpha cells, couples to MAP kinase cascade to mediate pheromone response; ligand bound receptors are endocytosed and recycled to the plasma membrane; GPC |
| TOM71 | Translocase of the Outer Mitochondrial membrane | Mitochondrial outer membrane protein with similarity to Tom70p; probable minor component of the TOM (translocase of outer membrane) complex responsible for recognition and import of mitochondrially directed proteins |

*continued on the next page*

**Genes in putative mating cluster 1** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| UBP9 | UBiquitin-specific Protease | Ubiquitin carboxyl-terminal hydrolase, ubiquitin-specific protease that cleaves ubiquitin-protein fusions |
| YCL065W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; overlaps HMLALPHA1 |
| YCR041W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data |
| YKL177W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene STE3 |
| YLR040C | N/A | Putative protein of unknown function; localizes to the cell wall; predicted to be a GPI-attached protein; upregulated by Mcm1p-Alpha1p transcription factor; partially overlaps the dubious ORF YLR041W; YLR040C is not essential |

Table A.6: **Genes in putative mating cluster 7 formed by JFuzzy-K from data of Brem *et al.* (2002).**

| Name | Short description | Full description |
|---|---|---|
| AGA2 | A-aglutinin | Adhesion subunit of a-agglutinin of a-cells, C-terminal sequence acts as a ligand for alpha-agglutinin (Sag1p) during agglutination, modified with O-linked oligomannosyl chains, linked to anchorage subunit Aga1p via two disulfide bonds |
| ASG7 | a-Specific Gene | Protein that regulates signaling from a G protein beta subunit Ste4p and its relocalization within the cell; specific to a-cells and induced by alpha-factor |
| BAR1 | BARrier to the alpha factor response | Aspartyl protease secreted into the periplasmic space of mating type a cells, helps cells find mating partners, cleaves and inactivates alpha factor allowing cells to recover from alpha-factor-induced cell cycle arrest |
| BST1 | Bypass of Sec Thirteen (Secretion) | GPI inositol deacylase of the ER that negatively regulates COPII vesicle formation, prevents production of vesicles with defective subunits, required for proper discrimination between resident ER proteins and Golgi-bound cargo molecule |
| HOP2 | HOmologous Pairing | Meiosis-specific protein that localizes to chromosomes, preventing synapsis between nonhomologous chromosomes and ensuring synapsis between homologs; complexes with Mnd1p to promote homolog pairing and meiotic double-strand break repair |
| LAG2 | Longevity Assurance Gene | Protein involved in determination of longevity; LAG2 gene is preferentially expressed in young cells; overexpression extends the mean and maximum life span of cells |
| MEC3 | Mitosis Entry Checkpoint | DNA damage and meiotic pachytene checkpoint protein; subunit of a heterotrimeric complex (Rad17p-Mec3p-Ddc1p) that forms a sliding clamp, loaded onto partial duplex DNA by a clamp loader complex; homolog of human and S. pombe Hus1 |
| MFA1 | Mating Factor A | Mating pheromone a-factor, made by a cells; interacts with alpha cells to induce cell cycle arrest and other responses leading to mating; biogenesis involves C-terminal modification, N-terminal proteolysis, and export; also encoded by MFA2 |
| MFA2 | Mating Factor A | Mating pheromone a-factor, made by a cells; interacts with alpha cells to induce cell cycle arrest and other responses leading to mating; biogenesis involves C-terminal modification, N-terminal proteolysis, and export; also encoded by MFA1 |
| STB5 | Sin Three Binding protein | Transcription factor, involved in regulating multidrug resistance and oxidative stress response; forms a heterodimer with Pdr1p; contains a Zn(II)2Cys6 zinc finger domain that interacts with a pleiotropic drug resistance element in vitro |
| STE2 | STErile | Receptor for alpha-factor pheromone; seven transmembrane-domain GPCR that interacts with both pheromone and a heterotrimeric G protein to initiate the signaling response that leads to mating between haploid a and alpha cells |

*continued on the next page*

**Genes in putative mating cluster 7** (*continued*)

| Name | Short description | Full description |
|---|---|---|
| STE6 | STErile | Plasma membrane ATP-binding cassette (ABC) transporter required for the export of a-factor, catalyzes ATP hydrolysis coupled to a-factor transport; contains 12 transmembrane domains and two ATP binding domains; expressed only in MATa cells |
| STE20 | STErile | Cdc42p-activated signal transducing kinase of the PAK (p21-activated kinase) family, involved in pheromone response, pseudohyphal/invasive growth, and vacuole inheritance; GBB motif (found in noncatalytic domains of PAKs) binds Ste4p |
| STE23 | STErile | Metalloprotease involved, with homolog Axl1p, in N-terminal processing of pro-a-factor to the mature form; member of the insulin-degrading enzyme family |
| STE24 | STErile | Highly conserved zinc metalloprotease that functions in two steps of a-factor maturation, C-terminal CAAX proteolysis and the first step of N-terminal proteolytic processing; contains multiple transmembrane spans |
| TFB3 | Transcription initiation Factor IIB | Subunit of TFIIH and nucleotide excision repair factor 3 complexes, involved in transcription initiation, required for nucleotide excision repair; ring finger protein similar to mammalian CAK and TFIIH subunit |
| YJL171C | N/A | GPI-anchored cell wall protein of unknown function; induced in response to cell wall damaging agents and by mutations in genes involved in cell wall biogenesis; sequence similarity to YBR162C/TOS1, a covalently bound cell wall protein |

Table A.7: **Genes in putative mating cluster 39 formed by JFuzzy-K from data of Brem *et al.* (2002).**

| Name | Short description | Full description |
|---|---|---|
| AKR2 | AnKyrin Repeat-containing protein | Also found in cluster 1 |
| ESS1 | ESSential | Peptidylprolyl-cis/trans-isomerase (PPIase) specific for phosphorylated serine and threonine residues N-terminal to proline; regulates phosphorylation of the RNA polymerase II large subunit (Rpo21p) C-terminal domain |
| HML$\alpha$1 | Hidden Mat Left $\alpha$ | Also found in cluster 1 |
| HML$\alpha$2 | Hidden Mat Left $\alpha$ | Also found in cluster 1 |
| LIF1 | Ligase Interacting Factor | Also found in cluster 1 |
| MAT$\alpha$1 | MATing type protein $\alpha$ | Also found in cluster 1 |
| MAT$\alpha$2 | MATing type protein $\alpha$ | Also found in cluster 1 |
| MF($\alpha$)1 | Mating Factor $\alpha$ | Also found in cluster 1 |
| MF($\alpha$)2 | Mating Factor $\alpha$ | Also found in cluster 1 |
| RSC6 | Remodel the Structure of Chromatin | Component of the RSC chromatin remodeling complex; essential for mitotic growth; homolog of SWI/SNF subunit Swp73p |
| SAG1 | Sexual AGglutination | Also found in cluster 1 |
| SEN54 | Splicing ENdonuclease | Subunit of the tRNA splicing endonuclease, which is composed of Sen2p, Sen15p, Sen34p, and Sen54p |
| STE3 | STErile | Also found in cluster 1 |
| THI22 | THIamine metabolism | Protein with similarity to hydroxymethylpyrimidine phosphate kinases; member of a gene family with THI20 and THI21; not required for thiamine biosynthesis |
| YCL065W | N/A | Also found in cluster 1 |
| YCR041W | N/A | Also found in cluster 1 |
| YJL150W | N/A | Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data |
| YKL177W | N/A | Also found in cluster 1 |
| YLR040C | N/A | Also found in cluster 1 |
| YLR041W | N/A | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the uncharacterized ORF YLR040C |
| YNL105W | Regulator of rDNA Transcription | Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps verified gene INP52; identified in a screen for mutants with decreased levels of rDNA transcription |
| YNL146W | N/A | Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the endoplasmic reticulum; YNL146W is not an essential gene |
| YRR1 | Yeast Reveromycin-A Resistant | Zn2-Cys6 zinc-finger transcription factor that activates genes involved in multidrug resistance; paralog of Yrm1p, acting on an overlapping set of target genes |
| YSR3 | Yeast Sphingolipid Resistance | Dihydrosphingosine 1-phosphate phosphatase, membrane protein involved in sphingolipid metabolism; has similarity to Lcb3p |