Abstract

ZOU, WEI. Transcriptional Regulatory Patterns in Yeast Revealed through Expression Quantitative Trait Locus Mapping. (Under the direction of Dr. Zhao-Bang Zeng.)

Expression Quantitative Trait Locus (eQTL) mapping combines large scale mRNA expression analysis and classical quantitative genetics methods to find the correlated variation between genome and transcriptome. We applied a sequential genome scan method to build genetic architectures for expression traits in a published yeast eQTL mapping data set. We compared mapping results from several variants of the method. We used the threshold controlling the genome-wise type I error rate to declare eQTLs, and used False Discovery Rate (FDR) to assess the overall error rate in the whole mapping study. In this mapping population, expression traits tend to be mapped onto its own sequence and avoid being mapped onto their transcriptional factors. Sequence variations around transcriptional factors are not preferentially associated with transcript abundance variation of their regulatory target genes, though the expression traits of transcriptional factors and their targets tend to have overlapping eQTL regions. Indirect trans-regulation mechanisms play a significant role in order to connect trans-acting eQTLs with certain biological pathways. We have developed a graphical tool to visualize the myriad relationships suggested by eQTL analysis. The tool allows dynamical superimposition of biological annotation onto declared eQTLs to find the matching cases between statistical patterns inferred from this mapping population and results of biological investigation on this organism. Finally, we proposed a probabilistic way of combining statistical patterns with biological annotations based on Bayes' Theorem. The resulting posterior probability that a gene in an eQTL causes the trait variation can be used to prioritize genes in an eQTL.

Transcriptional Regulatory Patterns in Yeast Revealed through Expression Quantitative Trait Locus Mapping

BY

WEI ZOU

A DISSERTATION SUBMITTED TO THE GRADUATE FACULTY OF

NORTH CAROLINA STATE UNIVERSITY

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BIOINFORMATICS

RALEIGH

2006

APPROVED BY:

TRUDY F.C. MACKAY	SUJIT K. GHOSH
ZHAO-BANG ZENG CHAIR OF ADVISORY COMMITTEE	BRUCE S. WEIR

To my parents

Biography

Wei Zou was born as the only child to his parents Heng Zou and Lipin Shen in Simao, Yunnan Province, China in October 1976. He was admitted to a bio-medical program jointly offered by Fudan University and Shanghai Second Medical University in 1994 and received a Bachelor of Medicine degree in July 1999. He then studied Human Genetics in the graduate school of Fudan University. In August 2002, Wei came to North Carolina State University (NCSU) to pursue a PhD degree in Bioinformatics. Under the direction of Dr. Zhao-Bang Zeng, Wei studied Quantitative Trait Locus Mapping.

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor Dr. Zhao-Bang Zeng for his thoughtful guidance and his unbelievable patience with me during the last several years. Without him, it would have been much more difficult for me to make progress and finish as scheduled. I have learned a lot from him both academically and personally. I feel very fortunate to have such an incredible advisor. I also thank Dr. Mackay, Dr. Ghosh and Dr. Weir for contributing their time to serve as the members of my advisory committee.

I really appreciate the help from the current and past members of Zhao-Bang's group: David Aylor, Sheng-Mao Chang and Jessica M. Maia. I enjoy the fruitful discussions with them a lot.

I also thank my friends in North Carolina State University Xiaoyi Gao, Xiaohua Gong, Jian Li, Sunil Suchindran, Jiaye Yu and Sihui Zhao for their friendship.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review

1.1 Introduction

Quantitative genetics studies the variation of quantitative traits. When Fisher laid down the basic theoretical foundations of quantitative genetics, the focus of study was to partition the overall variation into genetic and environmental ones. With the development of polymorphic markers for many species, current research interest is to partition genetic variation to individual quantitative trait loci (QTLs) in the genome as well as interaction among them (Zeng *et al.*, 1999). A QTL is a chromosomal region that is likely to contain genes that affect the phenotypic variation under study.

Given a mapping population, the fundamental information that any modern QTL mapping procedure relies on is chromosome recombination events in the population. Recombination creates small or large difference in the segregation pattern of DNA segments in the population. When the genome is properly tagged with genetic markers (section 1.2), QTL analysis discriminates DNA segments according to how their associated markers co-segregate with trait variation, and claims QTL when the association is significant in the mapping population (section 1.3).

In model-based QTL studies, markers are generally not attached with any genotypic effects. QTL can be anywhere between marker intervals. QTL position and effects are parameters to be estimated. In non-parametric approaches, markers are generally treated as explanatory variables for phenotypic variation. Linkage disequilibrium (LD) (section 1.4) between invisible causal genes for the trait and visible markers surrounding the causal genes attaches genetical meaning to these markers. Statistical details of QTL analysis are discussed in section 1.5. Section 1.6 is devoted to discuss thresholds for declaring QTL. Section 1.7 discusses a few attempts to go from a QTL to causal genes.

Recent research interest in systems biology makes transcriptome expression a subject of quantitative genetics study. Expression QTL (eQTL) analysis is the focus of this thesis. Section 1.8 gives the research outline on the topic.

1.2 Marker

Commonly used genetic markers are all DNA variations:

- restriction fragment length polymorphisms (RFLP)
- simple sequence repeats (SSR), or microsatellites
- variable number of tandem repeats (VNTR), or minisatellites
- single nucleotide polymorphisms (SNPs)

They have the following good properties: stably transmitted between generations, polymorphic in the population to differentiate inheritance patterns of chromosome regions, and there are efficient ways to measure marker genotypes.

SNPs have become a very popular kind of markers in the recent years for their superior genetic characteristics: abundantly available (about 1 in every 1200 base pair in human (`http://www.hapmap.org/`)); low mutation rate, much stabler than microsatellites. Recent advance in microarray technology, which greatly decreases the cost and increases the speed of genome-wise SNP genotyping (Wang *et al.*, 1998), is also an important factor for its wide application.

Markers are used to trace recombination events in the population. If significant LD only exists in a small genomic neighborhood around markers, as in natural populations, increasing marker density will increase the power and resolution to detect true underlying sequence variation for the trait. When there are only limited recombination events in the sample, dense markers will not be very helpful. Neighboring markers without any recombination event between them are equivalent with a single marker in detecting association. However, when a QTL is claimed around these non-recombinant markers, they are able to present more information about QTL boundaries than a single marker.

Marker density in a genomic region should be commensurate with the recombination rate in the region. Genotyping too many markers can waste money and time that may be used to increase the sample size. This is the rationale underlying the hierarchical genotyping procedure and tag SNPs selection procedure in HapMap project (The International HapMap Consortium, 2003).

1.2.1 Haplotype

A haplotype is a list of allelic types on a chromosome in their physical order. The simplest haplotype consists of two alleles at two loci. The most complex haplotype is the haploid genome.

It is always possible to combine individual markers into haplotypes and then detect association between the trait and the genomic region covered by the haplotypes. Haplotype-based analysis is motivated by biological observations like the case of APOE gene's effect on late-onset Alzheimer's disease: the disease susceptibility depends on the combination of alleles at two loci (Brouwer *et al.*, 1996). In association studies, it is not hard to find examples where there is a strong 'interaction effect' of two markers on the phenotype; but there is no significant 'main effect' for either marker (Culverhouse *et al.*, 2002; Hoh and Ott, 2003). Such association patterns will be hard to detect when markers are tested one by one, but easy for haplotype-based analysis. There have been many studies on the power comparison between single marker analysis and haplotype-based analysis, some of which favor the former analysis (LONG and LANGLEY, 1999) and some of which favor the later(AKEY *et al.*, 2001). NIELSEN *et al.* (2004) suggested that these contradicting findings can be explained by high order LD among two markers and one putative causal variant. Haplotype-based analysis is favored when there are moderate to high levels of three locus LD. To test markers one by one is a more powerful way when pairwise LD is high.

Haplotype frequencies capture the association pattern in a limited genomic region. The frequencies will be too low and highly variable for haplotypes covering an extended region. In order to detect joint effects of distant loci, or loci on different chromosomes, searching for interaction effects between individual loci is still an indispensable procedure to model quantitative traits.

1.3 Mapping population

Traditional QTL mapping experiment starts with two parental inbred lines, P_1 and P_2, derived from natural populations of the organism. Putative causal loci and informative markers are assumed to be homozygotic within a parental line, but fixed with different alleles between lines. Thus, phenotypic values of the trait of interesting, genetically affected by those causal loci, follow the same distribution within a line, and different distributions between lines. This is the basis of likelihood approaches for QTL mapping.

Certain markers, which are polymorphic in the nature population, can be homozygotic across P_1 and P_2. Such markers are considered as 'non-informative' for all mapping populations derived from P_1 and P_2, because their alleles do not correspond to a unique parental origin of their chromosomal regions. Similarly, there can be multiple types of alleles at each causal gene. P_1 and P_2 only contain two of them. This is one of the reasons that inferences and conclusions from a QTL analysis are generally restricted to the mapping population.

P_1 and P_2 can be crossed to produce F_1 population. Causal loci and informative markers become heterozygotic with an allele frequency of 0.5 for either type of allele. F_1 can be crossed with either parental line to create a backcross population, or with each other to create F_2. The later process can be repeated multiple times among relatives to create recombinant inbred lines (RIL).

Such well controlled cross designs are only available for animals and plants. In human, only observational studies are possible. Samples used for QTL mapping can be related individuals: a large number of nuclear families or several extended pedigrees; or unrelated ones from naturel population. Many pieces of information are missing in such samples compared with experimental populations: recombination events are not observed directly, but need to be inferred from pedigree structures; allelic frequencies are unknown and need to be estimated from the sample. When samples are from multiple families, there is much more genetic heterogeneity in forming the phenotype than that in inbred populations. Special attention is also needed to guard against population stratification.

1.4 Linkage disequilibrium (LD)

Two-locus LD is a measure of dependency between allele frequencies at two loci (Weir, 1996). A basic two-locus LD measure, 'D', is the difference between the two-locus haplotype frequency and the product of two allele frequencies at the two loci.

Mapping populations differ a lot as discussed in section 1.3, but they share one assumption in common: significant LD exists between putative causal genes and their approximal markers, which makes these markers segregate in a similar pattern as causal genes. Markers do not affect trait values by themselves genetically. But they can have effects in statistical modelling. Such an effect is a function of the effect of a causal gene, and LD between the causal gene and the marker.

In an experimental crossing population from inbred lines, LD between loci reach their maxima in F_1 populations, and expected to decrease in later sib-mating generations as more recombination events happen in meiosis. Parametric linkage analysis explicitly models the dependency of allele frequencies at marker loci and causal genes as a function of recombination rates among them (section 1.5).

For a natural population, when a new mutation was initially introduced, it was in perfect association with the haplotype on the chromosome where it was located. If it survived after generations of drift, it could propagate in the population. However, in the meanwhile, the recombination event would gradually erode the correlation between the mutant allele and alleles at neighboring loci on the original haplotype, unless certain allele combination on the haplotype was favored by selection as a whole (eg: high LD in MHC region of human genome due to balancing selection (Slatkin, 2000)). It would not take long before only those loci, which are very close in genetic distance with the mutant locus, could still preserve significant LD with the mutant. The assumption of a common historical mutation event underlying disease susceptibility for the general population is a key idea under whole-genome association study in human (Kruglyak, 1997). Similar arguments can be applied to linkage analysis in pedigrees: a mutation for disease susceptibility is introduced into the family through a founder.

Terwilliger *et al.* (1998) showed that excessive LD would exist between a disease predisposing allele and surrounding markers without a historical mutation or haplotype. They viewed a population as divided into two sub-populations with or without the disease allele; recombination was modelled as symmetric genome flow between two sub-populations; additional mutation at the disease locus as unidirectional gene flow; selection on the disease locus as sub-population size changes, etc. The elevated genetic drift in the smaller subpopulation carrying the disease allele will incur higher LD around the disease allele than other genomic area.

Besides recombination and drift, LD can be affected by the following effects.

- Selection. Selection is a constant force affecting allele frequencies given an environmental setting. If certain allele combinations from neighboring sites (eg: MHC locus) are favored by balancing selection, high and extended LD can be maintained in the population (Slatkin, 2000). When positive selection favors an allele of a single locus, due to 'hitchhiking effect', the haplotype where the allele is located will also be favored. The resulting increase in homogeneity of haplotypes in the chromosome region would increase LD surrounding the locus under positive selection.
- Population admixture. When two populations are mixed, LD will be

generated around loci with different allele frequencies in the two original populations.

- Inbreeding and associative mating, which make recombination less effective in eroding LD (Terwilliger *et al.*, 1998)
- Population expanding, which makes drift less effective in forming new LD in the population.

Affected by many random effects, LD itself is not so good a measurement of physical proximity between loci as recombination rate.

1.5 Statistical methods for QTL mapping

1.5.1 Single marker analysis

The basic principle of QTL mapping has been established in Sax (1923)'s work in beans. If there is co-segregation between the causal gene and a marker locus, mean values of the trait under study will differ among subject groups with different genotypes at the marker locus (Mackay, 2001a).

Such a principle leads immediately to a test procedure like student's t-test, where the difference between association study for natural populations and linkage analysis for experimental populations blurs.

However, in an experimental population, given a clear genetic map of markers, a formal likelihood approach can be applied (Zeng, 2000). Assuming normal error distribution of trait values, a likelihood ratio test can be performed to contrast the hypothesis that the recombination rate between the marker and a putative causal gene of the trait is 0.5 (unlinked), versus the hypothesis that the recombination rate is less then 0.5 (linked). Such an analysis reveals that the test statistic is not only affected by the map distance, but also by the genetic effects of putative causal genes.

1.5.2 Interval mapping (IM)

Lander and Botstein (1989) provided a likelihood based framework of IM to estimate causal gene locations within marker intervals. Compared with the single marker analysis, IM allows separate tests for QTL location and effects.

The likelihood function in IM is composed of two parts (Ott, 1999):

- conditional probability of a QTL genotype given neighboring marker genotypes, containing QTL location parameters;
- conditional probability of phenotype given a QTL genotype, containing QTL effect parameters. For binary traits (diseased or not), the notion of penetrance is used instead of genetic effects (eg: main effects, dominant effects) for continuous traits. Penetrance is the probability of showing the disease phenotype when a subject carries a disease allele. Continuous trait values are usually assumed to be random variables from normal distributions with means as linear functions of genetic parameters.

For a backcross population produced by crossing F_1 and P_1, for j-th individual ($j = 1..N$), let y_j be the trait value, Q_{ij} be the genotype of a putative causal locus in the genomic interval bounded by marker M_{ij} and $M_{(i+1)j}$; let 1 indicate heterozygosis at a locus and 0 indicate homozygosis. Table 1.1 shows $Pr(Q_{ij}|M_{ij}, M_{(i+1)j})$.

Assuming individuals with $Q_{ij} = 0$ have trait values sampled from $N(u, \sigma^2)$, and individuals with $Q_{ij} = 1$ have trait values from $N(u + b, \sigma^2)$, the likelihood function

Table 1.1: Conditional probabilities of QTL given its flanking markers in a backcross design

Marker genotype	Expected frequency	QTL genotype	
	of Marker Genotype	$Q_{ij} = 0$	$Q_{ij} = 1$
$M_{ij} = 0, M_{(i+1)j} = 0$	$(1 - r_0)/2$	1	0
$M_{ij} = 0, M_{(i+1)j} = 1$	$r_0/2$	$1 - \rho$	ρ
$M_{ij} = 1, M_{(i+1)j} = 0$	$r_0/2$	ρ	$1 - \rho$
$M_{ij} = 1, M_{(i+1)j} = 1$	$(1 - r_0)/2$	0	1

$\rho = r_1/r_0$, where r_0 is the recombination rate between M_{ij} and $m_{(i+1)j}$, and r_1 is the recombination rate between M_{ij} and Q_{ij}. Probability for double crossing-over is ignored.

that Q_{ij} is a causal locus for the trait is given by

$$\begin{aligned} L(u, b, \sigma) &= \prod_{j=1}^{N} [p_{0j}\phi(\frac{y_i - u}{\sigma}) + p_{1j}\phi(\frac{y_i - u - b}{\sigma})] \quad (1.1) \\ p_{0j} &= Pr(Q_{ij} = 0|M_{ij}, M_{(i+1)j}) \text{ (from table 1.1)} \\ p_{1j} &= Pr(Q_{ij} = 1|M_{ij}, M_{(i+1)j}) \text{ (from table 1.1)} \end{aligned}$$

where ϕ is the standard normal density function, b is the difference of mean trait values between individuals with $Q_{ij} = 1$ and $Q_{ij} = 0$. b is thus the genetic main effect of Q_{ij}. Maximal likelihood estimate (MLE) of r_1, the map distance between M_{ij} and Q_{ij}, is not obtained analytically, but estimated through a grid search in the genome. We can perform the following analysis at equally spaced positions (eg: 1cM apart, i.e., set $r_1 = 1cM, 2cM, ...$) within each marker interval. It is called genome scan in QTL mapping studies.

Expectation-maximization (EM) algorithm can be used to compute MLE for parameters in formula 1.1 after cycling through a series of expectation (E) and maximization (M) steps until convergence of estimates (Zeng, 2000). In the E-step of cycle t, posterior probability $Pr(Q_{ij} = 1|M_{ij}, M_{(i+1)j}, y_i)$, denoted as $P_j^{(t)}$, can be obtained as

$$P_j^{(t)} = \frac{p_{1j}\phi(\frac{y_i-u-b}{\sigma})}{p_{1j}\phi(\frac{y_i-u-b}{\sigma}) + p_{0j}\phi(\frac{y_i-u}{\sigma})}$$

In the M-step, derivatives of log likelihood function (formula 1.1) with respect to parameters are equated to zero to solve for MLE of cycle t.

$$\begin{aligned}
\frac{\partial \log L}{\partial b} &= \sum_{j=1}^{N} P_j^{(t)} \frac{y_j - u - b}{\sigma^2} \\
\frac{\partial \log L}{\partial u} &= \sum_{j=1}^{N} \frac{P_j^{(t)}(y_j - u - b) + (1 - P_j^{(t)})(y_j - u)}{\sigma^2} \\
\frac{\partial \log L}{\partial \sigma^2} &= \sum_{j=1}^{N} \frac{P_j^{(t)}(y_j - u - b)^2 + (1 - P_j^{(t)})(y_j - u)^2}{2\sigma^4} - \frac{N}{2\sigma^2}
\end{aligned}$$

In this likelihood approach, we can test the existence of a QTL r_1 away from M_{ij} using likelihood ratio statistic $LRS = -2\log[L(\tilde{u}, b = 0, \tilde{\sigma}^2)/L(\hat{u}, \hat{b}, \hat{\sigma}^2)]$. We could not compare the statistic to a χ_1^2 distribution to determine the significance, because the huge number of such tests are performed through the genome scan. We will discuss the issue of choosing a threshold to declare QTLs in section 1.6.

1.5.3 Composite interval mapping (CIM)

Given three consecutive markers x_{i-1}, x_i, x_{i+1}, when crossover interference can be overlooked, i.e., $Pr(x_{i+1}|x_i, x_{i-1}) = Pr(x_{i+1}|x_i)$, Zeng (1993) found that conditional on x_i,

- the covariance between x_{i-1} and x_{i+1} is zero,
- the expected partial regression coefficient of trait values on x_{i-1} will not be affected by possible QTLs between x_i and x_{i+1} or beyond x_{i+1}, if further ignoring epistasis.

Motivated by these findings, CIM is proposed to add flanking markers as covariates to the likelihood function as formula 1.1. As noted by Zeng (2000), CIM will reduce the chance of interference from nearby QTLs on hypothesis testing and parameter estimation for the current QTL. However, it will increase the variance of estimates when introducing correlated explanatory variables into the model.

In the CIM procedure implemented in QTL cartographer (Basten *et al.*, 1994), unlinked but 'important' markers are also included in the likelihood function to control genetic variation unexplained by the current QTL. Those 'important' markers are obtained by running a stepwise regression analysis for all markers across the genome. Multiple interval mapping extends such use of unlinked controlling markers into explicitly modelling quantitative trait architecture including multiple QTLs and epistasis.

1.5.4 Multiple interval mapping (MIM)

It has been suggested that a major QTL's effect should be considered when searching for secondary QTLs (Paterson *et al.*, 1988). Multiple interval mapping addresses this issue formally and suggests a model selection procedure to search for the best genetic model for the quantitative trait (Kao and Zeng, 1997; Kao *et al.*, 1999).

For m putative causal genes for the trait, the model of MIM is specified as (Zeng, 2000)

$$y_i = u + \sum_{r=1}^{m} \alpha_r x^*_{ir} + \sum_{r \neq s \subset (1,...,m)}^{t} \beta_{rs}(x^*_{ir}x^*_{is}) + e_i \tag{1.2}$$

where

- y_i is the phenotypic value of individual $i, i = 1, 2, ..., n$;
- u is the mean of the model;
- α_r is the main effect of r-th putative causal gene, $r = 1..m$;
- x^*_{ir} is an indicator variable denoting genotype of r-th putative causal gene, which follows a multinomial distribution with parameters similar to the conditional probabilities shown in table 1.1;
- β_{rs} is the possible epistatic effect between r-th and s-th putative causal gene, assuming there are t such effects ;
- e_i is an environmental effects assumed to be normally distributed

As shown by Kao and Zeng (1997); Kao *et al.* (1999), given a genetic model (number, location and interaction of multiple QTLs), this linear model suggests a likelihood function similar to formula 1.1 but more complex. EM algorithm can be used to maximize the likelihood and obtain MLE of parameters.

The following model selection method is used to transverse the genetic model space in QTL Cartographer (Basten *et al.*, 1994):

1. Forward selection of QTL main effects sequentially. In each cycle of selection, pick the best position of an additional QTL, and then perform a likelihood ratio

test for its main effect. If a test statistic exceeds the critical value, this effect is retained in the model. Stop when no more QTLs can be found.

2. Search for epistatic effects between QTL main effects included in the model, and perform likelihood ratio tests on them. If a test statistic exceeds the critical value, the epistatic effect is retained in the model. Repeat the process until no more significant epistatic effects are found.

3. Re-evaluate significance of each QTL main effect in the model. If the test statistic for a QTL falls below the significant threshold conditional on other retained effects, this QTL is removed from the model. However, if a QTL is involved in a significant epistatic effect with other QTL, it is not subject to this backward elimination process. This process is performed stepwisely until no effects can be dropped.

4. Optimize estimates of QTL positions based on the currently selected model. Instead of performing a multi-dimensional search around the regions of current estimates of QTL positions, estimates of QTL positions are updated in turn for each region. For the r-th QTL in the model, the region between its two neighbor QTLs is scanned to find the position that maximizes the likelihood (conditional on the current estimates of positions of other QTLs and QTL epistasis). This refinement process is repeated sequentially for each QTL position until there is no change on estimates of QTL positions. However, this optimization procedure is not used in my application because in this procedure, EM algorithm performs not so well and convergence is very slow.

An important issue in model selection is the significance level to include or eliminate effects. In regression analysis, such threshold is usually decided based on information criteria, which has the following general form

$$IC = -2(\log L_k - kc(n)/2) \tag{1.3}$$

where L_k is the likelihood of data (eg: formula 1.1) given a genetic model with k parameters, $c(n)$ can take following forms

- $c(n) = \log(n), (BIC)$
- $c(n) = 2$, (AIC)
- $c(n) = 2\log(\log(n))$
- $c(n) = 2\log(n)$
- $c(n) = 3\log(n)$

When the penalty for an additional parameter in the model is low, many main effects will be includes, and much more epistatic effects will be tested, which result in a very slow searching process. This is the case when $c(n) = 2\log(n)$. On the other hand, it is also a good idea to avoid high penalty like $c(n) = 3\log(n)$, in order to favor strong QTL interaction effects involving QTLs with less strong main effects.

In analyzing the yeast data, we would include a new QTL to the model if its corresponding LRS is larger than the 90% quantile of the empirical null distribution from permutation tests (section 1.6). That is equivalent to a penalty value between $2\log(n)$ and $3\log(n)$.

1.5.5 Model selection in dense marker set

As there are more and more markers genotyped in the mapping population, it seems less and less important to assume an unknown QTL locus between markers as in IM. On the contrary, given current dense marker set from SNPs arrays and limited mapping resolution, QTL loci are expected to straddle over many markers. 'Rudimental' QTL scanning procedures, similar to those used by Sax (1923), begin to appear more frequently in the literature. For example, non-parametric version of t-test, Wilcoxon-Mann-Whitney test, on single markers was applied to backcross population in yeast population (Brem *et al.*, 2002).

When the number of markers greatly exceeds the number of observations, and more than one markers are to be put into the model, QTL analysis turns out to be a model selection process: selecting a subset of markers to find the best way to combine them to explain the trait variation.

Using available methods, even two dimensional exhaustive search is a computational and statistical problem if we have thousands of markers across the genome (Culverhouse *et al.*, 2002; Hoh and Ott, 2003; Marchini *et al.*, 2005), and have thousands of traits, which is typical when doing expression QTL mapping (Brem *et al.*, 2002). Simulation studies, however, have shown that two dimensional search is more powerful than single locus scanning even when the underlying genetic model is additive one or with three way interaction (Marchini *et al.*, 2005).

Rather than exhaustive search of all marker combinations, sequential search of multiple loci for a trait, requires detectable main effect for the components of interaction effect (Hoh and Ott, 2003), though it is more powerful (Storey *et al.*, 2005). Data mining technique like decision tree, and pattern recognition might give alternative research algorithms, though it is not as popular as linear models in association study community now (Cook *et al.*, 2004; Shah and Kusiak, 2004; Storey *et al.*, 2005; Zhang and Bonney, 2000). Decision tree analysis, also known as an automatic interaction detection method (Hartigan, 1975), handles all explanatory variables as a component in certain interaction by default. High order interactions are detected in the same way as two way interactions without significant increase in computational complexity, though more samples are required.

1.5.6 Bayesian methods

Sen and Churchill (2001) proposed a Bayesian approach to sample QTL genotypes from their posterior distribution conditional upon marker genotypes and phenotypes, which offers an unified framework to handle various issues in QTL mapping: nonnormal and multivariate phenotypes, covariates, and genotyping errors. Their multiple imputation method differs from common Markov chain-Monte Carlo (MCMC) procedures in that a two step approach is used to get the posterior distribution: first, QTL genotypes are sampled only conditioning on marker genotypes; then, genotypes are weighted by the likelihood of phenotypes given the sampled QTL genotype. Although it is able to accommodate multiple QTLs and their interactions in the genetic model for a trait, current computational power limits the model complexity that this type of approach can handle in reasonable time.

1.5.7 LD mapping in natural populations

For natural populations, besides non-parametric methods based on contingency tables for case-control studies, or transmission/disequilibrium test (TDT) for families, in the work of Luo *et al.* (2000), an analogous likelihood function as formula 1.1 for linkage analysis was suggested. The difference between them is that, for natural populations, the joint probability of flanking marker genotypes and a putative causal gene genotype is a function of estimated marginal allele frequencies and LD coefficient between the loci, instead of recombination rates. A table, which is similar to table 1.1, is parameterized by LD.

1.5.8 Estimation of recombination rates from natural populations

Li and Stephens (2003) showed that recombination rates could be estimated from a random sample of a population while estimating haplotype frequencies based on Ewens' sampling theory (Ewens, 1972). Their algorithm was applied by Evans and Cardon (2005) to samples from several human populations. They found that estimate of recombination rates had a higher correlation coefficient across populations than that of LD measures. McVean *et al.* (2004) developed their bayesian approach based on coalescent theory and found that recombination rates estimated from random samples agreed well with those from pedigree data.

It has been known for a long time that LD coefficient, specially in a short range, is highly affected by random factors other than by recombination (section 1.4). Larger values of LD coefficient do not necessarily correspond to closer physical distance in both empirical and theoretical studies (Abecasis *et al.*, 2001; Pritchard and Przeworski, 2001). The irregular change of LD along chromosomes may pose a problem to parametric LD mapping in natural populations. It will also cause problems to locate disease genes by searching the genome around markers showing significant correlation with the phenotype according to non-parametric LD mapping results. On the other hand, recombination rate is a better candidate as a measure of distance. It is the unit of genetic maps. Although crossover interference will affect the additivity of recombination rates estimated from two small neighboring chromosomal intervals, it seems that it could only be a minor problem in pedigree data or experimental populations with a small number of generations. In modelling the evolution of natural populations, it is reasonable to view recombination events at each chromosomal section as independent poisson events (Li and Stephens, 2003) taking place at different generations of the population history.

It seems there is no theoretical difficulty in incorporating phenotypical data into the likelihood function for haplotype reconstruction suggested by Li and Stephens (2003) to extend their work into a QTL mapping algorithm:

- adding hypothetical QTL alleles to the haplotype to be reconstructed, which corresponds to the multi-locus approach of defining correlation between markers and a QTL (Jorde, 2000);
- connecting QTL genotype to phenotypic observations using penetrance functions (for binary traits) or assuming certain phenotypic value distribution functions conditional on QTL genotypes (usually normal distributions for quantitative traits).

It is possible that using MCMC algorithm to get the posterior distributions of putative QTL effects in each marker interval would bring huge computational burden.

1.6 Threshold to claim QTL

When one of the QTL mapping methods discussed in section 1.5 is applied, one important issue is how to choose a threshold to declare QTL. General asymptotic results for regression and likelihood ratio tests are not applicable in genome scans, given the large number of correlated tests and the limited sample size. Lander and Botstein (1989) discussed this issue for interval mapping with some elegant theoretical arguments.

1.6.1 Controlling type I error rates

When markers are dense and the sample size is lareg, Lander and Botstein (1989) showed that an appropriate threshold for LOD score (1 LOD = $2 \log 10$ logarithm of likelihood ratio statistic) is $(2 \log 10)t_\alpha$, where t_α solves the equation $t_\alpha = (C + 2Gt_\alpha)\chi^2(t_\alpha)$. C is the number of chromosomes of the organism. G is the length of the genetic map, measured in Morgans. $\chi^2(t_\alpha)$ is the probability that a random variable from a χ^2_1 distribution is less than t_α. Thus, it is not easy to get an analytic solution of t_α.

Although this threshold is derived under the assumption that the genome is completely covered by markers, their simulation studies showed that the threshold will decrease quite slowly with the increase of the genetic distance between markers. Such a conservative theoretical threshold will work in many applications.

Churchill and Doerge (1994) proposed a method based on permutation tests to find an empirical threshold specifically for a QTL mapping study. Data were shuffled by randomly pairing one individual's genotype with another's phenotype, in order to simulate the null hypothesis of no intrinsic relationship between genotypes and phenotypes. Thus, this method taks into account the sample size, the genome size of the organism under study, the genetic marker density, segregation ratio distortions, and missing data.

According to Churchill and Doerge (1994), the genome-wise threshold to control type I error rate for mapping a single trait can be found in the following procedure.

1. shuffle the data N times by randomly pairing trait values with genotypes;
2. obtain the maximum test statistic in each of N shuffled data, which are assumed to be sampled from a distribution F_M;
3. the $100(1-\alpha)$ percentile of F_M is the critical value.

Shuffle, or permute, is to break the original relationships in the data by randomly pairing observed phenotypes and genotypes to form new samples.

This permutation procedure is equivalent to the Bonferroni correction for multiple testing when test statistics are independent. Suppose there are n such statistics $t_i, i = 1..n$ from a null distribution F. $F_M(T)$, the distribution function of maximum of n statistics, can be expressed as $Pr(\max(t_i) < T) = F(T)^n$. When we find a threshold T, such that $Pr(\max(t_i) > T) = 1 - Pr(\max(t_i) < T) \leq \alpha$, it is the same as to require $1 - F(T)^n = 1 - (1 - Pr(t_i > T))^n \leq \alpha$, or $Pr(t_i > T) \leq \alpha/n$, the Bonferroni adjusted threshold. When test statistics are correlated, the permutation method discussed above provides an empirical estimate of F_M and imposes a stringent genome-wise type I error control.

A related permutation procedure was suggested by Doerge and Churchill (1996) for mapping procedures like MIM where QTLs are declared sequentially using a forward selection procedure. Two methods were suggested to find a genome-wise threshold for the second QTL while controlling effects of the first QTL.

- Conditional empirical threshold (CET). Mapping subjects are put into blocks according to the genotype of the marker identified as (or closest to) the first QTL. Permutation is applied within each block. Following the procedure described above by Churchill and Doerge (1994), maximal null statistics of each genome scan are collected and CET is obtained. One problem of CET is that markers linked to the first QTL will continue to show association with the trait variation as in the original data. To avoid CET being elevated by such markers, it is suggested to exclude the complete chromosome where the first QTL is located when collecting null statistics.
- Residual empirical threshold (RET). The residues from the genetic model with the first QTL are used as new phenotypic values to be permuted. Maximal null statistics from genome-wide scans are then collected to find RET.

It can be noticed from table 1 and 2 of Doerge and Churchill (1996)'s paper, the values of both CET and RET decrease in later cycles of the sequential genome scan. Thus, such thresholds provide more power to declare multiple QTL than the unconditional threshold.

1.6.2 False discovery rate (FDR)

In QTL mapping, as the number of markers grows, it is less likely that a casual variant is missed because it is not in LD with any marker. On the other hand, the number of markers showing significant correlation with the phenotype by chance is also expected to grow, when the type I error rate for each test is controlled at the same level. To handle this multiple testing issue, stringent family-wise control of type I error is usually applied, which is designed to control the probability of making at least one false discovery in a genome-wise test. However, a more powerful approach is to control false discovery rate (FDR) (Benjamini and Hochberg, 1995), or to control the expected proportion of false discoveries among all the markers passing a threshold, which is essentially to allow multiple false positive declarations when many 'significant' test statistics are found. Such a relaxation is driven by the nature of the problem under study: "It is now often up to the statistician to find as many interesting features in a data set as possible rather than test a very specific hypothesis on one item" (Storey, 2003).

Table 1.2: Possible outcomes from m hypothesis tests

	Accepted Null	Rejected Null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

According to the notation from Storey (2003), table 1.2 shows the possible outcomes when m hypotheses $H_1, H_2, ..., H_m$ are tested. For independent tests, Benjamini and Hochberg (1995) provided a procedure (known as linear step-up procedure or BH procedure) to control expected FDR, i.e., $E[\frac{V}{R}|R > 0] \times Pr(R > 0)$ at the desired level $\alpha\frac{m_0}{m}$ or α (since m_0 is generally unknown, a conservative up-bound estimate $m_0 = m$ is used) as following:

- sort P values from smallest to largest such that $P_{(1)} \leq P_{(2)}...P_{(m)}$;
- starting from $P_{(m)}$, compare $P_{(i)}$ with $\alpha\frac{i}{m}$
- let k be the first time $P_{(i)} \leq \alpha\frac{i}{m}$, reject all $P_{(1)}$ through $P_{(k)}$.

Benjamini and Yekutieli (2001) showed that "if test statistics are positively regression dependent on each hypothesis from the subset corresponding to true null hypotheses (PRDS), the BH procedure controls FDR at level $\alpha\frac{m_0}{m}$". For QTL mapping, PRDS can be interpreted as following (Sabatti *et al.*, 2003): if two markers have correlated allele frequencies and neither is related biologically to the trait, test statistics associated with the two markers should be positively correlated. Such positive correlation is intuitively correct and supported by simulation results (Sabatti *et al.*, 2003).

To check the performance of BH procedure on FDR control in genome-wise QTL scan for a single trait, Sabatti *et al.* (2003) considered a simulated case-control study in human. Three susceptibility genes are simulated to affect the disease status. They do not interact with each other and are located on different chromosomes. The results confirm that BH procedure does control the expected value of the FDR for single-trait genome-wise scan, although the actual FDR in a certain simulation replicate might exceed the rate that BH procedure tries to control. For multiple-trait QTL analysis, Benjamini and Yekutieli (2005) considered 8 positively or negatively correlated traits. Using simulation study, they showed BH approach works for multiple trait analysis too.

According to Benjamini and Yekutieli (2005), to control FDR for QTL analysis in each trait at level α does not always mean the overall FDR for these multiple traits is also α: if there are k independent and nonheritable traits, the overall FDR should be $1-(1-\alpha)^k \approx k\alpha$. It is safer to control FDR for all the tests simultaneously.

Yekutieli and Benjamini (1999) suggested to make use of dependency structure in data, rather than treat them as annoying cases. They expected an increase of testing power when using an empirical true null statistic distribution instead of assuming some theoretical ones. Empirical null distributions are used extensively in pFDR and local FDR as discussed in later sections of this chapter.

Though BH approach is a handy and intuitive tool supported by strict statistical theories, it should be used with caution.

- BH approach controls the expected value of FDR. Simulation studies showed that the actual FDR for a particular data set can be higher (Sabatti *et al.*, 2003).
- $E(FDR) = E[\frac{V}{R}|R>0]Pr(R>0)$ can be controlled at a designed level α by reducing $Pr(R>0)$ instead of $E[\frac{V}{R}|R>0]$ (Storey, 2002). Weller *et al.* (1998) are the first ones to apply FDR criteria in QTL mapping area. They claimed that 75% of the 10 QTLs declared in their study were true by controlling FDR at 25% using BH approach, Zaykin *et al.* (2000) pointed out that the interpretation was wrong because $E[\frac{V}{R}|R>0]$ could be much higher than $E(FDR)=25\%$ when $Pr(R>0)$ is small. It is $E[\frac{V}{R}|R>0]$, also known as pFDR discussed in the next section, that corresponds to the proportion of false discovery. Thus, when $\Pr(R=0)=1-\Pr(R>0)$ is high, or the number of positive findings from a test is low, FDR control is not appropriate because FDR is misleading in this situation.

- Fortunately, Weller (2000) could argue back by the fact that $\Pr(R > 0)$ in their case should be close to 1: given $R = 10$ and assuming R follows a Poisson distribution, $\Pr(R = 0)$ is very small. In later FDR literature, the assumption that $\Pr(R > 0) \approx 1$ is widely adopted (Benjamini and Yekutieli, 2005; Storey and Tibshirani, 2003).

1.6.3 positive discovery rate (pFDR)

pFDR, or $E[\frac{V}{R}|R > 0]$ was initially advocated by Storey (2002, 2003). Benjamini and Hochberg (1995) considered pFDR in their 1995 paper, but preferred FDR mainly because of its property brought about by the flexibility of $\Pr(R > 0)$. One can not determine an arbitrary threshold $\alpha, \alpha < 1$ and guarantee that pFDR $\leq \alpha$ regardless of the actual proportion of true null hypothesis in all tests. When $\frac{m_0}{m} = 1$, pFDR should be fixed at 1 and cannot be controlled at α. In this case, however, FDR=pFDR$\times \Pr(R > 0)$. Its value can be lowered to α by reducing the rejection region and pushing $Pr(R > 0)$ towards α.

Given a traditional type I error control procedure and conditional on $R > 0$, pFDR works naturally for estimation (instead of control) of the proportion of true/false discovery in R positive findings. Storey (2002) presented a bayesian interpretation of pFDR, and an estimation procedure for tests based on P values. Assuming there are m tests $H_1, H_2, ..., H_m$ with associated P values $P_1, P_2, ..., P_3$, and

- $H_i = 0$ denoting the i-th null hypothesis is true, and $H_i = 1$ otherwise;
- to model the uncertainty in each hypothesis test, each H_i is viewed as an identical and independent random variable from a Bernoulli distribution with $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = \pi_1 = 1 - \pi_0$;

- P_i follows a distribution f_0 if $H_i = 0$, and f_1 if $H_i = 1$; follows a mixed distribution unconditionally: $f_m = \pi_0 f_0 + \pi_1 f_1$;
- $\Gamma = [0, \gamma]$ is the common rejection region for all H_i;

then

$$\begin{aligned} pFDR(\gamma) &= \Pr(H_i = 0|P_i \in \Gamma) && (1.4) \\ &= \frac{\Pr(H_i = 0)\Pr(P \leq \gamma|H_i = 0)}{\Pr(P \leq \gamma)\Pr(R > 0)} && (1.5) \end{aligned}$$

where $\Pr(H_i = 0) = \pi_0$; $\Pr(P \leq \gamma|H_i = 0) = \gamma$, which is the pre-defined type I error rate; $\Pr(P \leq \gamma)$ can be estimated by $\frac{R}{m}$; $\Pr(R > 0)$ can be estimated by its lower bound $1 - (1 - \gamma)^m$. $\Pr(R > 0)$ equals to its lower bound when the power of the test $\Pr(P \in \Gamma|H = 1) = \gamma$.

Define $A = [\lambda, 1]$ as an 'accept region', i.e., assume $Pr(H = 0|P \in A) \approx 1$, Efron *et al.* (2001) gave an upper bound of π_0:

$$\frac{\int_A f_m(z)dz}{\int_A f_0(z)dz} = \frac{\int_A [\pi_0 f_0(z) + (1 - \pi_0)f_1(z)]dz}{\int_A f_0(z)dz} \geq \frac{\int_A \pi_0 f_0(z)dz}{\int_A f_0(z)dz} = \pi_0 \quad (1.6)$$

Applying formula 1.6 to this case, Storey and Tibshirani (2003) suggested that

$$\pi_0 = \frac{\#\{P > \lambda\}/m}{(1 - \lambda)} \quad (1.7)$$

Thus a conservative estimator of pFDR is

$$p\widehat{FDR}(\gamma) = \frac{\#\{P > \lambda\}\gamma}{R(1 - \lambda)[1 - (1 - \gamma)]^m} \quad (1.8)$$

To remove arbitrariness in choosing λ, two methods have been proposed.

- Bootstrap method (Storey, 2002): form bootstrap samples of $P_1, P_2, ...P_m$, find a λ through grid search in the interval $(0, 1)$ that minimizes the variability of pFDR estimation.
- Extrapolation method (Storey and Tibshirani, 2003): use a natural cubic spline smoothing method to find the upper bound of π_0 when $\lambda \to 1$, or the accept region A approaches an empty set.

Benjamini and Yekutieli (2005) stated explicitly the following: "pFDR is not a pFDR-controlling testing procedure. It is capable only of estimating the pFDR once a fixed rejection threshold is being used...". Storey (2002), however, proposed q value, a pFDR analogue of P value, to find a rejection region Γ given a specific pFDR rate. q value is the minimal pFDR when rejecting a hypothesis with P value P_i, or $min_{t \geq P_i} pFDR(t)$.

Storey and Tibshirani (2003) hypothesized that the above procedures would result in conservative estimates of pFDR and q values when there is weak dependence among test statistics. For gene expression traits, "because genes behave dependently in small groups (i.e., pathways), with each group essentially being independent of the others", the weak dependence condition will hold.

1.6.4 local FDR

The Bayesian interpretation of pFDR extends naturally to local FDR (Efron *et al.*, 2001), denoted as fdr in this paper:

$$\begin{aligned} fdr(T_i) &= \Pr(H_i = 0|T_i) \qquad (1.9) \\ &= \frac{\pi_0 f_0}{\pi_0 f_0 + \pi_1 f_1} = \frac{\pi_0 f_0}{f_m} \qquad (1.10) \end{aligned}$$

where T_i is the test statistic associated with H_i; $H_i = 0$ denoting the i-th null hypothesis is true, and $H_i = 1$ otherwise; T_i follows a distribution f_0 if $H_i = 0$, and f_1 if $H_i = 1$; follows a mixed distribution unconditionally: $f_m = \pi_0 f_0 + \pi_1 f_1$.

There is great similarity between formula 1.4 and formula 1.9, Efron (2005) showed that q value associated with T_i is equivalent with $E_{t \geq T_i} fdr(t)$. Such a connection is used by Storey *et al.* (2005) to control FDR in expression QTL mapping (his formula 9, 10 and 11).

The key part in estimating fdr is to estimate $\frac{f_0}{f_m}$. It is possible to assume certain standard distribution as f_0 and estimate f_m directly with non-parametric regression (Efron, 2005). The following is an alternative method (Efron *et al.*, 2001; Storey *et al.*, 2005) that is used extensively in my thesis work:

1. permute data under null hypothesis B times, and obtain test statistics $z_{ij}, i = 1..m, j = 1..B$;

2. estimate $\frac{f_0}{f_m}$ from T_i and z_{ij} (see below);

3. estimate π_0 using formula 1.7.

$\frac{f_0}{f_m}$ is estimated in the following way:

1. pool all T_i and z_{ij} into bins;

2. create an indicate variable y, let $y = 1$ for each T_i and $y = 0$ for each z_{ij}. Thus, $\Pr(y = 1) = \frac{f_m}{f_m + Bf_0}$ in each bin;

3. obtain a smooth estimate of $\widehat{\Pr}(y = 1)$ in each bin from an overall regression curve across bins, by combining natural cubic spline with generalized linear models for binomially distributed response variables;

4. equate $\widehat{\Pr}(y = 1)$ with $\frac{f_m}{f_m + Bf_0}$, and get a moment estimate of $\frac{f_0}{f_m}$.

It can be noticed that H_i and its associated T_i or P_i are assumed to be from a mixture distribution in both pFDR and fdr estimation. Thus, as pointed out by Storey (2003), there is a connection between multiple hypothesis testing and classification. For each test, the test procedure is to classify H_i as 0 or 1, or accepted or rejected. Classification decisions can be made based upon T_i, with a rejection region Γ: if $T_i \in \Gamma$, we classify H_i as 1. In chapter 4, we use such connection to classify a gene in an eQTL as 'affecting the trait' or not.

1.7 From QTL to gene

Most of QTLs identified through a linkage study will be 3-10 cM wide (Mackay, 2001b). LD study was originally proposed as a fine mapping procedure following a linkage study to further reduce the possible regions where candidate genes might be located (Snell *et al.*, 1989). With the recombination events accumulated in generations, LD study was considered to be able to locate causal genes with a resolution

$\leq$ 1 cM in the best cases (DE LA CHAPELLE and WRIGHT, 1998). It is possible to confirm or eliminate candidate genes one by one experimentally using deficiency and complementation mapping (MACKAY, 2001b), though it is an expensive and slow process.

In recent expression QTL (eQTL) analysis, where the traits of interest are the complete transcriptome, HUBNER *et al.* (2005) noticed the overlapping of physical locations between

- pQTLs (p for physiological traits) for spontaneously hypertension in SHR rat, and
- eQTLs detected in a recombinant inbreed strain obtained by crossing SHR and BN strain.

For example *Cd36* is believed to contribute to hypertension with its null mutation (AITMAN *et al.*, 1999) in SHR. A strong cis-acting eQTL for *Cd36* transcript abundance, detected in both fat and kidney tissues of the mapping population, was found within previously detected hypertension pQTL. Similarly, BYSTRYKH *et al.* (2005) presented 8 genes as candidate genes for hematopoietic stem cell (HSC) turnover. All 8 genes had a cis-acting eQTL in known pQTLs for HSC turnover. And 3 of them had already been implicated with the phenotype.

It seems that a major reason for the enthusiasm on relating pQTL with cis-acting eQTL comes from the stable statistical properties of cis-acting eQTLs and clear biological interpretation for pQTL (BYSTRYKH *et al.*, 2005; HUBNER *et al.*, 2005):

- when raising the significant level of claiming an eQTL, the percentage of cis-acting eQTL increases gradually to 100%;
- cis-acting eQTLs were shared between different tissues and organisms;

- cis-acting eQTL can demonstrate a large effect so that the inherence mode of an expression trait behaves like a Mendelian one;

It should be noticed that when declaring an eQTL as cis-acting, it is the same as to reduce the possible location of genetic determinants for an expression trait from a cis-acting eQTL to an even smaller genomic region containing only the gene itself, since it is very unlikely that a cis element for a gene is located in another gene. It should also be noticed that when declaring a cis-acting eQTL, genomic location information of the transcript is used. This would remind us to use the comprehensive gene annotation information organized in Gene Ontology (Ashburner *et al.*, 2000), or other biological knowledge to prioritize genes in an eQTL.

1.8 Outline of research for expression QTL mapping

The classical traits in quantitative genetics are morphological ones. QTL mapping is to search genetic factors for them, with the knowledge that these traits are the results of gene expression and protein interaction controlled by these genetic factors. Systemical study of gene expression has become possible with the microarray-based gene expression assay. Better understanding of the genetic architecture for the transcriptome, an intermediate layer in Central Dogma, will not only tell us the overall picture of transcriptional control, but can also associate morphological variations with certain genes' transcription profiles (Schadt *et al.*, 2003) and the biochemical pathways that these genes participate in. See table 1 by Gibson and Weir (2005) for summary of eQTL studies performed as far.

In Chapter 2, eQTL data in yeast published by Brem and Kruglyak (2005) were reanalyzed using MIM. Permutation tests were performed to find genome-wise thresholds for individual QTL. pFDR was further estimated to assess false discoveries in the rejection regions. Consistency and inconsistency with previous analysis were investigated. In order to find biological explanations for declared eQTLs, available genome annotation for yeast was superimposed onto statistical patterns, based upon gene name comparison. Chapter 3 describes a web-based software to visualize eQTL distribution patterns with known gene networks overlayed. Chapter 4 presents a Bayesian approach to prioritize genes in an eQTL according to their posterior probabilities of affecting the trait. Chapter 5 suggests future works on eQTL mapping.

1.9 References

Abecasis, G., E. Noguchi, A. Heinzmann, J. Traherne, S. Bhattacharyya, N. Leaves, G. Anderson, Y. Zhang, N. Lench, A. Carey, L. Cardon, M. Moffatt, and W. Cookson, 2001 Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* **68**: 191–197.

Aitman, T. J., A. M. Glazier, C. A. Wallace, L. D. Cooper, P. J. Norsworthy, F. N. Wahid, K. M. Al-Majali, P. M. Trembling, C. J. Mann, C. C. Shoulders, D. Graf, E. S. Lezin, T. W. Kurtz, V. Kren, M. Pravenec, A. Ibrahimi, N. A. Abumrad, L. W. Stanton, and J. Scott, 1999 Identification of Cd36 (Fat) as an insulin-resistance gene. *Nat Genet* **21**: 76–83.

AKEY, J., L. JIN, and M. XIONG, 2001 Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.* **9**: 291–300.

ASHBURNER, M., C. BALL, J. BLAKE, D. BOTSTEIN, H. BUTLER, J. CHERRY, A. DAVIS, K. DOLINSKI, S. DWIGHT, J. EPPIG, M. HARRIS, D. HILL, L. ISSEL-TARVER, A. KASARSKIS, S. LEWIS, J. MATESE, J. RICHARDSON, M. RINGWALD, G. RUBIN, and G. SHERLOCK, 2000 Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

BASTEN, C., B. WEIR, and Z. ZENG, 1994 Zmap-a QTL cartographer. In *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software*, volume 22, edited by C, S., G. JS, B. B, J. CHESNAIS, W. FAIRFULL, J. GIBSON, B. KENNEDY, and E. BURNSIDE, 5th World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada, 65–66.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Roy. Stat. Soc. B.* **57**: 289–300.

BENJAMINI, Y., and D. YEKUTIELI, 2001 The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**: 1165–1188.

BENJAMINI, Y., and D. YEKUTIELI, 2005 Quantitative Trait Loci Analysis Using the False Discovery Rate. *Genetics* **171**: 783–790.

BREM, R., G. YVERT, R. CLINTON, and L. KRUGLYAK, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

Brouwer, D. A. J., J. J. van Doormaal, and F. A. J. Muskiet, 1996 Clinical chemistry of common apolipoprotein e isoforms. *J Chromatogr B Biomed Appl.* **678**: 23–41.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–71.

Cook, N., R. Zee, and P. Ridker, 2004 Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med.* **23**: 1439–53.

Culverhouse, R., B. K. Suarez, J. Lin, and T. Reich, 2002 A Perspective on Epistasis: Limits of Models Displaying No Main Effect. *Am. J. Hum. Genet.* **70**: 461–471.

de la Chapelle, A., and F. A. Wright, 1998 Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *PNAS* **95**: 12416–12423.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285–94.

Efron, B., 2005 Local False Discovery Rates. http://www-stat.stanford.edu/ brad/papers/False.pdf.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher, 2001 Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**: 1151–1160.

Evans, D., and L. Cardon, 2005 A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet.* **76**: 681–7.

Ewens, W., 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**: 87–112.

Gibson, G., and B. Weir, 2005 The quantitative genetics of transcription. *Trends in Genetics* **21**: 616–623.

Hartigan, J., 1975 *Clustering Algorithms.* John Wiley and Sons, New York.

Hoh, J., and J. Ott, 2003 Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics* **4**: 701–709.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman, 2005 Integrated transcriptional profiling and linkage analysis for identificat ion of genes underlying disease. *Nat Genet* **37**: 243–253.

Jorde, L., 2000 Linkage Disequilibrium and the Search for Complex Disease Genes. *Genome Res.* **10**: 1435–1444.

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**: 653–65.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* **152**: 1203–1216.

Kruglyak, L., 1997 What is significant in whole-genome linkage disequilibrium studies? *Am J Hum Genet.* **61**: 810–2.

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

Li, N., and M. Stephens, 2003 Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**: 2213–2233.

Long, A. D., and C. H. Langley, 1999 The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genome Res.* **9**: 720–731.

Luo, Z. W., S. H. Tao, and Z.-B. Zeng, 2000 Inferring Linkage Disequilibrium Between a Polymorphic Marker Locus and a Trait Locus in Natural Populations. *Genetics* **156**: 457–467.

Mackay, T. F., 2001a The genetic architecture of quantitative traits. *Annu Rev Genet* **35**: 303–39.

Mackay, T. F., 2001b Quantitative trait loci in Drosophila. *Nat Rev Genet* **2**: 11–20.

Marchini, J., P. Donnelly, and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**: 413–417.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004 The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science* **304**: 581–584.

Nielsen, D. M., M. G. Ehm, D. V. Zaykin, and B. S. Weir, 2004 Effect of Two- and Three-Locus Linkage Disequilibrium on the Power to Detect Marker/Phenotype Associations. *Genetics* **168**: 1029–1040.

Ott, J., 1999 *Analysis of human genetic linkage.* 3rd edition, the Johns Hopkins University Press, Baltimore,Maryland.

Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln, and S. D. Tanksley, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**: 721–6.

Pritchard, J., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. *Am J Hum Genet.* **Jul**: 1–14.

Sabatti, C., S. Service, and N. Freimer, 2003 False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders. *Genetics* **164**: 829–833.

Sax, K., 1923 THE ASSOCIATION OF SIZE DIFFERENCES WITH SEED-COAT PATTERN AND PIGMENTATION IN PHASEOLUS VULGARIS. *Genetics* **8**: 552–560.

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.

Sen, S., and G. A. Churchill, 2001 A Statistical Framework for Quantitative Trait Mapping. *Genetics* **159**: 371–387.

Shah, S. C., and A. Kusiak, 2004 Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine* **31**: 183–196.

Slatkin, M., 2000 Balancing Selection at Closely Linked, Overdominant Loci in a Finite Population. *Genetics* **154**: 1367–1378.

Snell, R., L. Lazarou, S. Youngman, O. Quarrell, J. Wasmuth, D. Shaw, and P. Harper, 1989 Linkage disequilibrium in Huntington's disease: an improved localisation for the gene. *J Med Genet.* **26**: 673–5.

Storey, J. D., 2002 A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64**: 479–498.

Storey, J. D., 2003 the Positive False Discovery Rate: a Bayesian Interpretation and the q-Value. *Annals of Statistics* **31**: 2013–2035.

Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440–5.

Terwilliger, J., S. Zllner, M. Laan, and S. PSSbo, 1998 Mapping Genes through the Use of Linkage Disequilibrium Generated by Gene tic Drift: 'Drift Mapping' in Small Populations with No Demographic Expansion. *Hum Hered* **48**: 138–154.

The International HapMap Consortium, 2003 The International HapMap Project. *Nature* **426**: 789–96.

Wang, D. G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander, 1998 Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**: 1077–1082.

Weir, B., 1996 *Genetic data danalysis II.* sinauer.

Weller, J. I., 2000 Using the False Discovery Rate Approach in the Genetic Dissection of Complex Traits: A Response to Zaykin et al. *Genetics* **154**: 1919–.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin, and M. Ron, 1998 A New Approach to the Problem of Multiple Comparisons in the Genetic Dissection of Complex Traits. *Genetics* **150**: 1699–1706.

Yekutieli, D., and Y. Benjamini, 1999 Resampling-based false discovery rate controlling procedure for dependent test statistics. *J. Stat. Plan. Inference* **82**: 171–196.

Zaykin, D. V., S. S. Young, and P. H. Westfall, 2000 Using the False Discovery Rate Approach in the Genetic Dissection of Complex Traits: A Response to Weller et al. *Genetics* **154**: 1917–1918.

Zeng, Z., 1993 Theoretical Basis For Separation Of Multiple Linked Gene Effects In Mapping Quantitative Trait Loci. *Proc Natl Acad Sci U S A* **90**: 10972–10976.

Zeng, Z., C. Kao, and C. Basten, 1999 Estimating the genetic architecture of quantitative traits. *Genet Res.* **74**: 279–289.

Zeng, Z.-B., 2000 Statistical Methods for Mapping Quantitative Trait Loci. unpublished lecture notes.

Zhang, H., and G. Bonney, 2000 Use of classification trees for association studies. *Genetic Epidemiology* **19**: 323–332.

Chapter 2

Multiple Interval Mapping of Expression Traits in Yeast

2.1 INTRODUCTION

Transcriptional control is one of the most important steps for an organism to express the genetic information stored in its sequence as well as to respond to environmental changes (Ihmels *et al.*, 2004). Recent advance of genomic technology has made it possible to quantify transcript abundance systematically, as well as to genotype genetic markers covering the whole genome in a segregating population. Provided with these tools, expression QTL (eQTL) analysis has been applied to study inheritance of thousands of similar traits in hope to find general rules of genetic control of transcriptional regulation (Brem *et al.*, 2002; Bystrykh *et al.*, 2005; Chesler *et al.*, 2005; Schadt *et al.*, 2003).

On the other hand, the basic idea of comparing trait means among groups in the mapping population with different allelic types at a genetic locus, could be traced back to Sax (1923). This idea of directly analyzing markers one by one was implemented as Wilcoxon-Mann-Whitney Test by Brem *et al.* (2002); Yvert *et al.* (2003) and as single marker analysis by Bing and Hoeschele (2005), and applied to the yeast data published by Brem *et al.* (2002); Yvert *et al.* (2003).

Brem and Kruglyak (2005) re-analyzed the data, focusing on genetic architecture of expression traits and suggesting that the majority of heritable traits are controlled by multiple eQTLs. Storey *et al.* (2005) presented a sequential search algorithm and declared 170 2-QTL models for the expression traits in the same data while controlling false discovery rate (FDR) at 10%. In this paper, we applied multiple interval mapping (MIM) analysis (Kao *et al.*, 1999) to the data once again. The idea of MIM is to simultaneously fit multiple main and epistatic genetic effects in a model to integrate searching QTLs with inferring genetic architecture of quantitative traits (Zeng *et al.*, 1999). With MIM, 1242 traits were found with at least 2 eQTLs while overall FDR is estimated as 8%.

QTL mapping result is a summary of correlation patterns between expression phenotypes and genotypes, while biological annotation of yeast genome is a summary of biological investigation of yeast for years. Our bioinformatics analysis of the mapping result reveals that the correspondence between the two kinds of summaries not only provides probable biological explanations for the detected statistical correlation; but also suggests transcription control as the result of mixing two kinds of genetical information in this mapping population. As a key step of biological activity, transcriptional control affects the whole genome's mRNA expression through genetically coded transcriptional regulatory networks, or the direct trans-acting effect. On the other hand, the network activity is modulated by genetic variations across the genome through the indirect trans-acting mechanism, where the effect is passed along biochemical pathways and affects genes with shared biological functions. Based on our mapping result and with careful interpretation, a much stronger signal of direct trans-acting regulation by transcriptional factors was detected, as compared with Yvert *et al.* (2003). However, indirect transcription control was still important to account for most of trans-acting eQTLs observed. On one hand, for about 28% eQTLs, we could find that at least one gene within an eQTL interacts with the trait gene according to current knowledge about yeast biology. On the other hand, 50 groups of trait genes were found to share both eQTL and Gene Ontology (GO) (Ashburner *et al.*, 2000) annotations, while the annotation was further shared by a gene in the shared eQTL region. We developed scalable vector graphics (SVG) to dynamically display all types of superimposed annotation on the background of eQTL mapping result (`http://statgen.ncsu.edu/~wzou/svg/svg2.htm`).

2.2 METHODS

The data contains 6195 unique expression traits of 112 haploid segregants obtained by crossing laboratory (BY) and wild (RM) strains of budding yeast (Brem *et al.*, 2002; Brem and Kruglyak, 2005; Yvert *et al.*, 2003). 2956 genetic markers were genotyped with about 1000 markers showing at least one recombinant .

2.2.1 QTL Mapping

Log ratio of expression signals were normalized according to the two step procedure by Wolfinger *et al.* (2001) and standardized before supplied to MIM procedure in QTL Cartographer (Basten *et al.*, 2002).

Likelihood ratio tests (LRT) have been playing a key role in QTL mapping since the introduction of interval mapping (Lander and Botstein, 1989). A QTL would be declared if a locus is significantly more likely to carry the causal genetic variation than otherwise. In MIM, multiple QTLs could be included into a genetic model in a series of genome scans. In each cycle of genome scan, LRT were performed on and between genetic markers conditional upon QTLs found in previous cycles. Maximal likelihood ratio statistics (LRS) in one pass of genome scan for a trait will give rise to a QTL if it is above certain threshold. For each trait, we permuted the data 20 times and applied one pass of genome scan for each permuted data. Maximal LRS from each permutation was pooled to form the null distribution of unconditioned LRS. The 90% quantile of the distribution was used as the threshold. The choice of 90%

quantile was to favor possible interaction effects with minor main effects (Marchini *et al.*, 2005).

Because of the limitation of computational power, we could not get conditional null distributions for maximal LRS in the 2-nd cycle or later genome scans. We assumed these distributions should be similar as the unconditional one. Furthermore, according to the conditional empirical threshold estimation (CET) procedure (Doerge and Churchill, 1996), markers around previously declared QTLs should be excluded in the conditional genome scan, which would reduce the number of genome-wise tests and lower the maximal null LRS as well as the threshold. To test the idea, we ran regression analysis between trait values and marker genotypes on permuted data, where F statistics associated with markers are asymptotically equivalent with LRS. The null distribution of genome-wise maximal F statistics for unconditional scans is with mean 11.16 and standard deviation 2.86. We followed the CET procedure (Doerge and Churchill, 1996) to obtain the null distribution of maximal F statistics conditional upon the marker showing strongest association with the trait. The distribution looks very similar to the unconditional one (Figure 2.1), with mean 11.14 and standard deviation 2.86. Thus, compared with the unconditional null distribution, conditional null distributions would skew a little bit towards small values of statistics. As the result, using the unconditional threshold for conditional scans will lead to a conservative inference.

Only main effects were considered above. We used the same threshold to add interaction effects into genetic models of traits with at least two QTLs. After that, backward elimination processes were applied to insure all declared genetic factors were statistically significant.

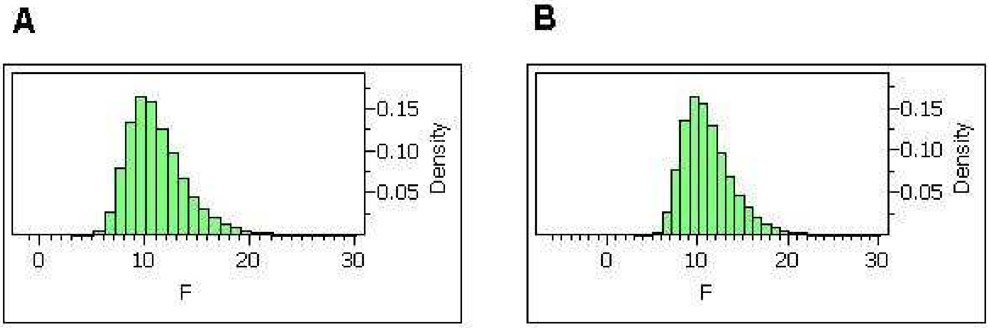

Figure 2.1: Null distributions of F statistics. A: The null distribution of genome-wise maximal F statistics for the first eQTLs. B: The conditional null distribution of genome-wise maximal F statistics for the second eQTLs, given the existence of the first eQTLs, obtained using CET.

After the model selection process, a LRS profile surrounding each QTL, conditional upon all the other significant genetic effects in the model, was obtained to get interval estimation of QTLs. For those QTLs with LRS peak above 15.4 (95% quantile of the null distribution), a 1.5 LOD support interval around a peak was declared as the QTL region.

2.2.2 FDR in Sequential Genome Scan

Empirical threshold discussed above is supposed to control genome-wise type I error for each eQTL. To find the overall errors for all the traits, we obtained empirical Bayesian estimate of the probability being a true eQTL for each eQTL according its LRS (Efron and Tibshirani, 2002; Efron *et al.*, 2001). For a list of maximal LRS obtained in one cycle of genome scan, the prior probability that a LRS in the list is associated with a false eQTL was calculated according to Efron *et al.* (2001). More specifically, all LRS that are smaller than the lower quartile of the list were considered to be purely generated by false eQTLs, resulting in a conservative estimation of the prior probability and hence FDR (Storey and Tibshirani, 2003). Let Q_{ij} the j-th eQTL of trait i; $Q_{ij} = 1$ denote the eQTL is true. Then, $\Pr(Q_{ij} = 1|LRS)$, the posterior probability that Q_{ij} is true given its LRS was obtained using a non-parametric logistic regression described by Efron *et al.* (2001); Storey *et al.* (2005). Thus, the probability that the multiple eQTL model for trait i is correct can be expressed as $\prod_j \Pr(Q_{ij} = 1|LRS)$. FDR for sequential genome scan is the average rate of declaring a false genetic model (Storey *et al.*, 2005):

$$FDR = \frac{1 - \sum_i \prod_j \Pr(Q_{ij} = 1|LRS)}{\text{number of traits}} \tag{2.1}$$

2.2.3 From eQTLs to Gene Lists

Biological interpretation of eQTL mapping result using bioinformatics approach is expected to be fruitful for *Saccharomyces cerevisiae*, one of the most studied organisms. For this purpose, we obtained the gene lists for all detected eQTLs. SNPs with known position in both the genetic and physical maps serve as anchors to translate eQTL genetic map intervals into sections of chromosome with physical boundaries. Lists of genes located in the intervals were then produced. Open Reading Frames (ORF) in these eQTL intervals are called 'pORF' in this paper since they are close to DNA polymorphism co-segregating with traits. Those ORFs, whose expression levels were used as traits, are called 'eORF' hereafter. We focused on ORFs since all trait genes are ORFs and most of non-ORF genes are not well annotated in the GO system which makes further analysis difficult. Mapping results are visualized in a two dimensional plot (Fig 2.2), where the 16 chromosomes of yeast are concatenated to form the horizontal and vertical axis; each eQTL is plotted as a small horizontal bar, whose vertical coordinate corresponds to the eORF and two horizontal coordinates are determined by the physical positions of the first and last pORFs of the eQTL. Thus, cis-acting eQTLs would be along the diagonal, while trans-acting eQTLs are not.

2.2.4 Generation of Random Mapping Results

Random mapping results were simulated to assess the probabilities that the observed eQTL distribution patterns could appear out of random if the same number eQTLs are declared. In the simulation,

- number of eQTLs for each eORF was generated according to the empirical distribution of number of eQTLs per trait in the real mapping result;
- the chromosome where an eQTL would be located was picked with the probability proportional to the chromosomal length;
- eQTL position was then picked from the chromosome randomly;
- length of the eQTL was generated according to a gamma distribution with shape parameter 1.96 and scale parameter 25659.8. These two values were obtained by fitting the versatile gamma density function to the histogram of eQTL physical length from the real mapping result.

100000 hypothetical mapping results were simulated in this way.

2.2.5 Indirect trans-acting effects between functionally related genes

When a pORF, as a transcriptional factor, was found within an eQTL of an eORF, as a corresponding target of the transcriptional factor, the eQTL was considered as supportive for the direct trans-acting mechanism and the effect of the transcriptional factor would be a probable biological explanation for the eQTL. However, such a mechanism is quite rare in this mapping population according to Yvert *et al.* (2003) and the results reported below. Thus, indirect trans-acting effects among genes participating similar biological processes may be needed as an explanatory biological mechanism for the off-diagonal eQTLs in Figure 2.2. To systemically detect eQTL distribution patterns which were compatible with current knowledge of gene function in yeast, off-diagonal bars were interpreted as two different types of lists: one was a horizontal list of pORFs in an eQTL; the other was a vertical list of eORFs whose eQTLs contain a common pORF. In a horizontal list (or an eQTL), we tried to identify a pORF that was related with the corresponding eORF using binary gene relationship stored in various yeast gene interaction datasets. For a vertical list of eORFs, we tested whether a similar annotation was shared by both the underlying pORF and multiple eORFs in the list, and tested whether such multiplicity could happen in a random collection of eORFs. In this step, unitary gene annotation for each gene was used. This approach was similar to finding a shared function for a subset of eORFs in a group of eORFs mapped to the same genomic bin by Brem *et al.* (2002), but without subjectively dividing the genome into bins and with an attempt to find an underlying pORF to explain the functional clustering. We consider this orthogonal dissection of mapping results very helpful in applying bioinformatics analysis following eQTL analysis.

It is straightforward to apply binary information between gene pairs onto those horizontal lists. To utilize unitary annotation for vertical lists, GO annotation for each eORF in a list was supplied to GO:TermFinder package (`http://search.cpan.org/~sherlock/GO-TermFinder-0.7/`) to detect over-represented GO terms in it. The package detects a subgroup of eORFs sharing the same GO term and tests whether the size of the subgroup could be expected by chance. GO terms for an eORF subgroup passing the significant test are called over-represented terms. We applied an additional requirement that those over-represented GO terms should match GO terms of the pORF in the shared eQTL region in 'cellular component'('C'), 'molecular function' ('F') and 'biological process'('P') ontology simultaneously. Statistical significance associated with matching in three ontologies guarded against random coincidence and allowed more confidence in further biological interpretation.

2.3 RESULTS

2.3.1 Mapping Result

5182 eQTLs for 3367 eORFs were detected by MIM. Table 2.1 shows the detailed process of sequential search. Very few interaction cases were retained: 49 interacting pairs for 38 expression traits. Overall FDR was estimated at less than 0.08. Average 1.5 LOD support interval of eQTLs is 50kb, containing about 27.7 ORFs in each interval. Their distribution across the genome is shown in Figure 2.2. Among the detected eQTLs, excluding those associated with 2 trait genes without location information, 737 eQTLs cover the physical locations of the corresponding eORFs, indicating the possibility of cis-regulation. These eQTLs are referred to as 'cis-acting eQTLs' in the following text. 409 eQTLs are on the same chromosomes of the corresponding eORFs, but do not overlap with the eORFs. The rest 4033 eQTLs are on different chromosomes from their eORFs.

Table 2.1: Sequential Genome Scan Results

cycle	#scanned[1]	#found[2]	#retained[3]
1	6195	3367	3354
2	3367	1617	1242
3	1617	578	422
4	578	197	122
5	197	66	37
6	66	10	5

[1] Number of traits for which a genome scanning was performed in the cycle
[2] Number of traits with one eQTL found in the cycle, controlling type I error rate at 10%. This threshold was liberal, in order to favor eQTLs involved in strong interaction effects but with main effects below a stricter threshold.
[3] Number of traits with one eQTL finally retained in the cycle, controlling type I error rate at 5% to declare each eQTL

The number of cis-acting eQTLs is much larger than that obtained under the random mapping scheme: where the number is normally distributed (verified through normal Q-Q plot) with mean 196 and standard deviation 14. Thus, just as discovered previously (Bing and Hoeschele, 2005; Chesler *et al.*, 2005) and shown in Figure 2.2 , 'cis-acting' is a major non-random pattern of expression regulation. In transcriptional regulation, cis-acting mechanism relies on cis-acting elements: DNA segments within or around the structural portions of a gene that interact with trans-acting factors in controlling its expression. The large number of cis-acting eQTLs in the mapping population suggests that the cis-acting mechanism could be a simple, unified piece of explanation for all such eQTLs. We will have more discussion about this pattern later.

2.3.2 Trans-regulation through Transcriptional Factor

The biological effect of a transcriptional factor (TF) on its regulatory target (TG) can cause the statistical correlation between expression variation of the TG and sequence variation around the TF. In yeast, 3417 genes are such TGs at certain physiological conditions (Luscombe *et al.*, 2004). In this mapping population, expression profiles of 1886 TGs are mapped to at least one eQTL. However, there are only 49 cases, where a TG's eQTL overlaps with the DNA sequence of its corresponding TF (Figure 2.2). These 49 TF-TG pair-wise relationship include the 3 TGs mapped to their common TF YJL206C, which is the only TF regulation case reported by Yvert *et al.* (2003).

Under the null hypothesis that eQTLs are generated randomly, the probability of observing 49 or less TGs mapped onto their TFs is close to zero. The null distribution is normal (verified through normal Q-Q plot) with mean 225 and standard deviation 16. Thus, contrary to cis-acting eQTLs, there is a clear statistical tendency for TGs to avoid being mapped onto their TFs.

Table 2.2: Transcriptional factor-target pairs supported by eQTL mapping

Target (eORF)	Transcriptional Factor (pORF)	Distance (bp)[1]	Correlation[2]
YDR214W	YGL073W	NA	0.01736359
YGR289C	YGR288W	NA	0.34216383
YPL032C	YIL131C	NA	0.10612629
YIR031C	YIR023W	NA	0.16205202
YPL275W	YJL206C	NA	0.07768735
YBL038W	YKL112W	NA	0.07175296
YJL026W	YLR176C	NA	-0.1139548
YPR141C	YMR070W	NA	0.05703913
YDL210W	YML027W	182727	-0.00229373
YGL089C	YCR040W	0	0.81213402 ⋆
YKL178C	YCR040W	0	0.91238004 ⋆
YPL187W	YCR040W	0	0.89949137 ⋆
YGL035C	YGL035C	0	1
YKL007W	YJR060W	0	0.42612154
YHR008C	YLR256W	0	-0.38249603
YML054C	YLR256W	0	-0.21111789
YOR065W	YLR256W	0	-0.48566585
YNL216W	YNL216W	0	1
YER001W	YOL089C	0	-0.29943447
YNL113W	YBR182C	-1	-0.22808445
YDR132C	YCR065W	-1	-0.32850968

Table 2.2 (continued)

Target (eORF)	Transcriptional Factor (pORF)	Distance (bp)	Correlation
YNL087W	YDR259C	-1	-0.3886451
YDR224C	YER111C	-1	0.12128574
YDR500C	YER111C	-1	0.11894158
YIR020W-A	YER111C	-1	-0.16581042
YKR042W	YER111C	-1	-0.13435156
YMR136W	YER111C	-1	0.21488443
YOL012C	YER111C	-1	0.10482232
YOR153W	YGL013C	-1	0.12909517
YNL007C	YGL073W	-1	0.02660671
YOR298C-A	YGL073W	-1	0.0200029
YBR040W	YHR006W	-1	0.03004001
YKL175W	YJL056C	-1	0.80896981 ⋆
YLR130C	YJL056C	-1	-0.28244773
YNL254C	YJL056C	-1	0.69537845 ⋆
YOR388C	YJL206C	-1	-0.00526971
YPL276W	YJL206C	-1	0.02872369
YPR191W	YKL109W	-1	0.59426705
YGR180C	YLR176C	-1	-0.17621602
YIL066C	YLR176C	-1	-0.00360481
YGR250C	YLR403W	-1	0.31036973
YLR109W	YML007W	-1	-0.1577697
YPL265W	YML099C	-1	0.20270183

Table 2.2 (continued)

Target (eORF)	Transcriptional Factor (pORF)	Distance (bp)	Correlation
YBL033C	YNL068C	-1	-0.24002106
YBR037C	YNL068C	-1	-0.30796576
YDR452W	YNL068C	-1	-0.28456703
YER078C	YNL068C	-1	-0.09968361
YLR439W	YNL068C	-1	-0.07303374
YPR035W	YNL068C	-1	0.23687292

[1] The closest distance between an eQTL of a pORF (also a target of a transcriptional factor in this table) and the pORF sequence itself. 'NA' means no eQTL was found for the pORF. 0 means there is overlapping between the trans-acting eQTL of the target (the eQTL contains the corresponding transcriptional factor) and cis-acting eQTL of the transcriptional factor, i.e. suggesting the secondary TG eQTL scenario. '-1' means no eQTL of the pORF was on the same chromosome as the pORF sequence.

⋆ These TF-TG pairs are among the 143296 ORF pairs with most correlation in expression, i.e, the absolute value of correlation coefficient is larger than 0.65.

[2] Pearson correlation of expression abundance between the target and the transcriptional factor.

The direct trans-acting mechanism is also difficult to detect by expression profile analysis. 47 known TF-TG pairs could be found in the 143296 most correlated gene pairs in terms of their expression levels (143296 is the total number of pORFs in all the eQTLs). This comparison shows that by implicating a certain number (143296 here) of gene pairs, eQTL mapping has the similar efficiency in including true TF-TG pairs as expression profile analysis. However, only 5 TF-TG pairwise relationships are shared by the two sets. Thus, relying on a different type of information than co-expression studies, eQTL mapping is able to capture some unique signal in transcriptional variation.

TGs are mapped to their TFs for different reasons. In these 49 pairs, the trans-acting eQTLs of 2 TGs (YGL035C and YNL216W) could be regarded as cis-acting eQTLs as well since these two TGs are TFs of themselves. In another 8 cases, the trans-acting eQTLs for the TGs overlap with the cis-acting eQTLs for the TFs. This coincidence brings some ambiguity for biological interpretation. One possible explanation is that there are two underlying genetic factors, one affecting expression variation of the TG, and another for the TF. An alternative explanation is that there is a single genetic factor affecting the expression of the TF only. The TF's abundance change could then possibly cause the TG's expression variation. In this case, there could still be correlation between the TG's expression and the genetic factor's sequence variation, which directly affects the TF's expression, and hence the TG will have an eQTL around the genetic factor. This is referred to as the 'secondary TG eQTL scenario' later on. Under this scenario, high correlation of transcript abundance between the TF-TG pair is expected. In the 8 pairs, it is found that the 3 TGs that are mapped onto transcriptional factor YCR040W are highly correlated at the expression level with their common TF (with Pearson correlation coefficient larger than 0.8). However, the correlations for the rest of 5 pairs are not always positive and high, suggesting heterogeneity of transcriptional regulation (Table 2.2).

More candidates for the 'secondary TG eQTL scenario' could be found if we look into the overlapping eQTLs for all the known TF-TG pairs. That is, we do not require the genomic regions to contain corresponding TFs but collect all eQTL distribution patterns that are compatible with the scenario.

318 TF-TG pairs (excluding possible self-regulation ones) were found with at least one overlapping eQTLs. Correlation coefficients of their transcript abundance had a mean of 0.13 (corresponding roughly to the 75% quantile of all pairwise correlation) and a median of 0.19 (82% quantile of all correlations) . If only considering 180 pairs with positive correlation, the mean and median were 0.4(96% quantile of all correlations). See Figure 2.3 for the histogram of these 318 correlations.

Using the same random mapping simulation procedure, the null distribution of the number of TF-TG pairs sharing eQTLs was found to be normal (verified through normal Q-Q plot) with mean 184 and standard deviation 35. Thus, it is unlikely that these eQTL-sharing cases were all generated through a random eQTL assigning process for TFs and TGs separately and independently. Possible explanations include that the TF-TG pairs have the tendency to be mapped together, or the 'secondary TG eQTL scenario', which is more convincing and parsimonious. In this scenario, a trans-acting eQTL affects the expression variation of a TF. Current knowledge about the transcriptional regulatory network in yeast could not relate the TF with any of genes in the eQTL. It is possible that there is a unknown TF in the eQTL; or that the eQTL represents a link in some feedback loops to fine-tune the activity of transcriptional regulation networks, or that the underlying mechanism for the eQTL is the variation in a signal transduction process which coordinates the TF's activity with intracellular and extracellular changes.

Thus, our mapping results could suggest at most 2.6% (49 of 1886) of known TGs with at least one eQTL were regulated through sequence variation around their corresponding TFs. Secondary TG eQTL scenario is able to suggest somewhat stronger and significant signals of trans regulation of TG expression through TF. Still only a small portion of trans-acting eQTLs could be attached with some biological explanations, if we restrict ourselves in a direct trans-acting relationship between TFs and TGs.

We offer explanation for the deficiency of statistical patterns reflecting the biological effect of TFs, and the abundance of statistical patterns for cis-acting mechanisms as follows. 1) The mapping population is composed of yeast cells in various stages of their cell cycles. Most TFs are only active during a certain phase of cell cycle (LUSCOMBE *et al.*, 2004). There could be internal inconsistency of active regulatory network topology among yeast segregants. Such inconsistency would lower the statistical power to detect co-segregating patterns between expression traits and polymorphic sites at the candidate loci. In this sense, a mapping population under certain external stimuli is expected to represent more faithfully the underlying regulation by TFs in response to stimuli. On the other hand, sequence variability in cis-acting elements can affect the basal transcriptional level of nearby genes in a consistent way across cell cycles. Thus, it is much easier to catch cis-acting mechanisms in the population. 2) TF might be too critical to accumulate sequence variation in normal laboratory strains or wild strains. They are more likely to exert large phenotypic effect and push the organism out of the phenotypic threshold of being normal. On the other hand, the direct influence from a cis-acting genetic polymorphism is always limited in its neighborhood. Purifying selection pressure is expected to be much stronger on transcriptional factors than on their targets. 3) We could not rule out the possibility that MIM procedure might have missed some epistatic eQTLs which could have supported more roles played by transcriptional factors. However, the complex biochemical interactions in transcriptional regulation networks would not guarantee complex statistical interactions (Barton and Keightley, 2002). All gene products act together with other genes' product in a cell, directly or indirectly, however, most of them have 'main effects' attributed to their own, and a lot of them have no detectable 'interaction effect' statistically.

2.3.3 Horizontal list analysis: biological relationships compatible with eQTLs

pORFs, which were functionally related to corresponding eORFs, were picked out using the following sets of information. That is, we collected matching cases between an eQTL, as a statistical pattern, and a biological relationship between a gene pair identified outside the current mapping study. Many biological relationships discussed below are not transcriptional regulations. But each of them has the potential to result in an indirect trans-acting effect between the gene pair. On the other hand, even an observation that an eQTL contains a transcriptional factor for the trait gene does not pinpoint the transcriptional factor as the causal gene for the eQTL. Each matching case suggests a plausible explanation for the eQTL, which should be the first hypothesis to be tested during in-depth study of the eQTL. 1) FinalNet (Lee *et al.*, 2004), which implements the idea of probabilistic view of gene relatedness by summarizing heterogeneous knowledge about yeast: mRNA co-expression across 497 microarrays, protein interaction predicted from genomic context of protein sequence (gene fusion and gene co-occurrence across organisms), functional relatedness inferred from literature co-citation and protein interaction experiments(mass spectrometry analysis for co-precipitated protein complexes, high throughput yeast two hybrid assays, and high-throughput synthetic lethal screens). Within each evidence, log likelihood score (LLS) for a pair of genes was calculated to represent the likelihood of functional linkage between the pair given the evidence. The finally reported LLS is a weighed sum of LLS scores from different lines of evidence. Only gene pairs with LLS ≥ 1.5 were used in the following comparison. 2) Three sets of annotation from Saccharomyces Genome Database (`http://www.yeastgenome.org/`, accessed Feb 19,2005) (Balakrishnan *et al.*, 2005): Complex GO-Slim, genes in macromolecular complexes; synthetic lethal, gene pairs put together because double mutation of the pair will be lethal to yeast; two hybrid, protein pairs physically interact with each other confirmed through yeast two hybridization assay. 3) MIPS (Mewes *et al.*, 2002), another protein complex database. 4) Transcriptional factor-target pairs (Luscombe *et al.*, 2004) as discussed previously. Figure 2.2 marks out matching cases between eQTLs and part of annotations listed above (check `http://statgen.ncsu.edu/~wzou/svg/svg2.htm` to dynamically view all matching patterns). Table 2.3 lists the number of matching cases with each annotation.

In all types of annotation information, finalNet offers the most matching cases: 747 pairs of eORF-pORF can be found in finalNet with LLS ≥ 1.5. This is expected since finalNet is based upon various sources of information. Assuming the number of matching cases is with a hyper-geometric distribution under the null hypothesis that these matching pairs appear by chance, the probability of observing 747 or more pairs is zero. Thus, eORF-pORF relationships obtained from eQTL analysis concentrate functionally related gene pairs in a statistically significant way. Indirect transcriptional modification between these related pairs from finalNet is a likely mechanism underlying certain trans-acting eQTLs. However, gene pairs which are related in finalNet do not overlap well with those in other kinds of evidence, which with specific focus and more concrete experimental support. The idea of building 'bottom-up' relationship network of gene functions (Fraser and Marcotte, 2004) has a lot of potential. Certainly there is room to improve its algorithm and to include more pieces of evidence including eQTL mapping results.

In all the 5182 declared eQTLs, 1442 have some connection with the biological evidences listed in table 2.3, or are compatible with the cis-acting mechanism. For the rest of eQTLs, they might be statistical artifact, but it is much more likely that they reveal novel biological connections between eORFs and pORFs. eQTL mapping should be treated as an important genome annotation tool in post-genomic era.

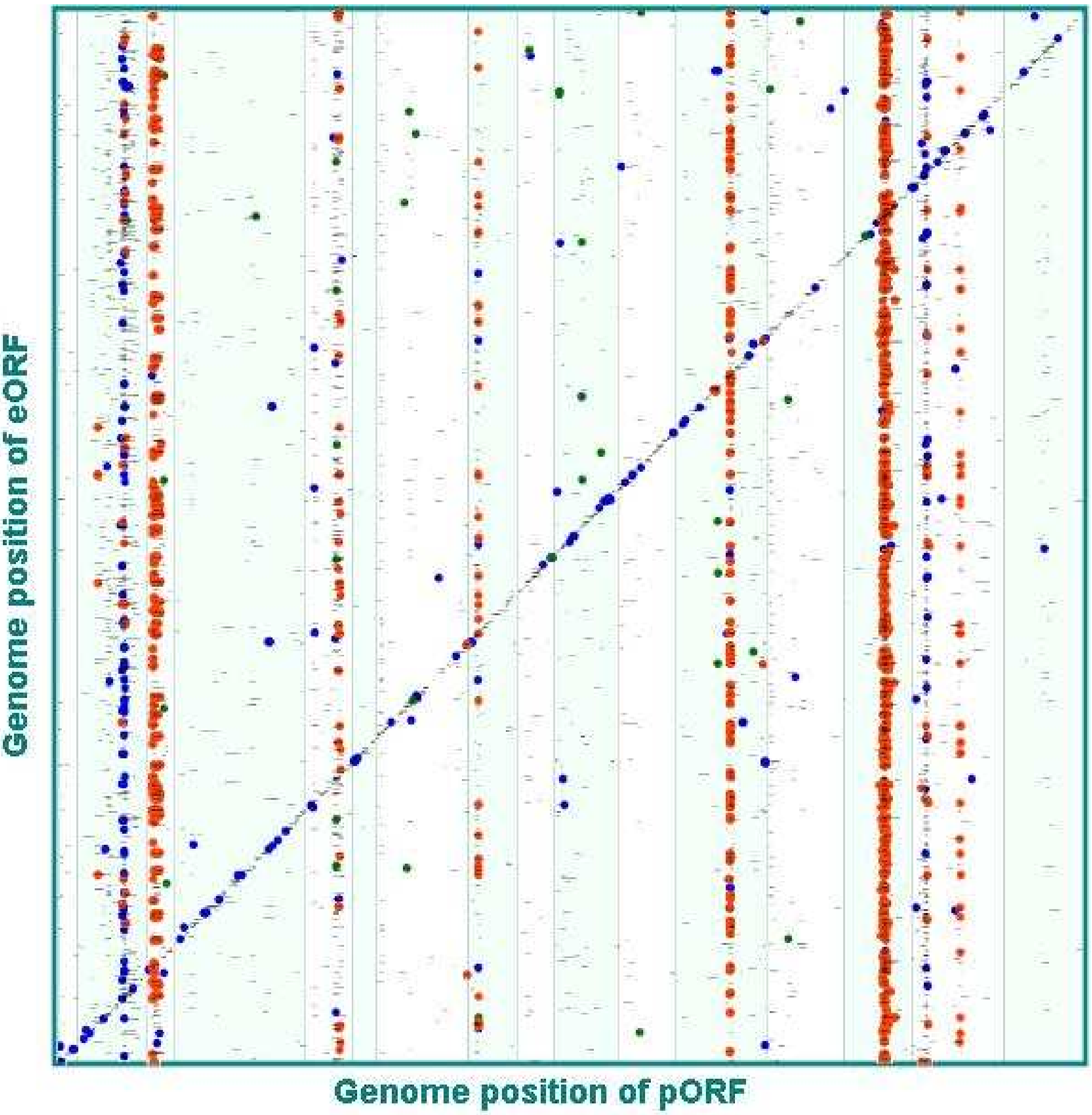

Figure 2.2: Two dimensional plot of mapping results. Red dots: 50 vertical lists of eORFs with over-represented GO terms matching GO terms of the gene in their shared eQTL region. In each vertical list, only eORFs annotated with the over-represented GO terms are marked. Blue dots: eORF-pORF relationship found in finalNet. Green dots: pORF-eORF is a known transcriptional factor-target pair. This picture is a screenshot from our SVG viewer, where users are allowed to zoom and pan the 2 dimensional plot to check detailed information about certain eQTLs, or to view different types of matching patterns as their request.

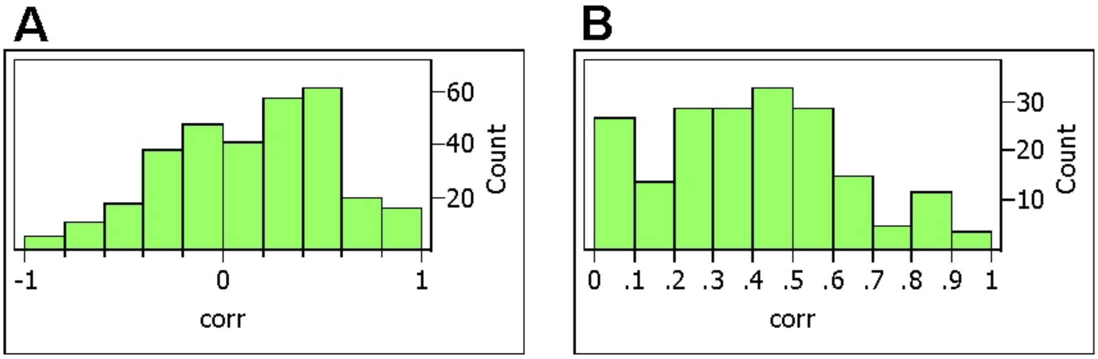

Figure 2.3: Histogram of correlation coefficients of transcript abundance between transcriptional factor and its target which have overlapping eQTLs. (A) For All 318 TF-TG pairs. Mean of correlation is 0.13. (B) 197 positively correlated TF-TG pairs only. Mean of correlation is 0.40

Table 2.3: Mapping results confirmed through various sources

evidence	#trait[1]	#eQTL[2]	#match[3]	#shared[4]	#known[5]	#genes[6]
finalNet	553	592	747	NA	72693	5349
Complex GO-Slim	227	277	734	74	117390	1242
MIPS	48	50	51	31	8250	871
Synthetic Lethal	29	30	34	0	4118	1048
Two hybrid	15	15	15	4	2007	1367
TF regulation	49	49	49	4	7073	3456

[1] number of traits with at least one pORF, where the eORF-pORF relationships are supported by the evidence

[2] number of eQTLs containing at least one pORF, which is related with the corresponding eORF according to the evidence

[3] total number of pORFs that are related with the trait genes according to the evidence

[4] number of eORF-pORF relationships supported by both the current evidence and by finalNet.

[5] total number of known pairwise relation in evidence

[6] total number of genes involved in the evidence

2.3.4 Vertical list analysis: shared biological functions and shared eQTLs among multiple eORFs

There are more obvious ways to detect such patterns other than using GO. For example, we tried to find whether gene pairs in the same MIPS protein complex or the same KEGG pathway (Kanehisa *et al.*, 2004) tend to share eQTLs. Most positive finding is related with mitochondrion. For example, 29 of 32 genes in mitochondrial ribosomal large subunit have at least one eQTL mapped in the genome. All of them have one eQTL in the middle of chromosome 14, overlapping with each other. In F0/F1 ATP synthase (located at mitochondrial inner membrane), 14 member genes have at lease one eQTL. 81 pairs of eORFs have shared eQTL regions. Besides using MIPS database, according to KEGG, eQTLs of eORFs in pathway 'Oxidative phosphorylation' clearly gather at Chromosome 15 (`http://statgen.ncsu.edu/~wzou/svg/yeastExample/group.svg`).

Sophisticated analysis detected more patterns of eQTL clustering. 50 full matching cases were detected between over-represented GO terms in vertical lists of eORFs and GO terms of pORF in all three GO ontologies. Such a fully matching means that the subgroups of eORFs annotated with over-represented GO terms and the underlying pORFs work at similar cellular component, with similar function and in similar biological process. See Table (2.4) for more information about the 50 lists.

Table 2.4: Over-represented GO terms detected in a vertical list which match GO terms of the pORF underlying the vertical list

pORF[1]	#eORF[2]	prob C[3]	prob F[4]	prob P[5]	#matching[6]
YBR001C	29	1.22×10^{-8}	1.07×10^{-7}	4.73×10^{-8}	5
YBR003W	29	1.22×10^{-8}	1.07×10^{-7}	4.73×10^{-8}	5
YBR142W	189	7.29×10^{-5}	1.93×10^{-2}	1.35×10^{-3}	5
YBR154C	442	1.73×10^{-2}	3.21×10^{-2}	1.35×10^{-7}	9
YBR167C	154	5.45×10^{-6}	3.25×10^{-2}	9.71×10^{-3}	12
YCL004W	162	1.08×10^{-6}	1.39×10^{-2}	7.29×10^{-5}	27
YCL005W-A	174	1.23×10^{-6}	7.03×10^{-3}	8.55×10^{-4}	59
YCL009C	174	1.23×10^{-6}	7.03×10^{-3}	8.55×10^{-4}	59
YCL017C	234	8.77×10^{-9}	1.15×10^{-3}	1.29×10^{-8}	97
YCL018W	234	8.77×10^{-9}	1.15×10^{-3}	1.29×10^{-8}	97
YCL019W	236	5.75×10^{-9}	1.25×10^{-3}	1.39×10^{-8}	98
YCL024W	261	1.37×10^{-9}	4.64×10^{-2}	$4.96\times10^{-1}0$	110
YCL030C	126	1.15×10^{-2}	4.93×10^{-2}	4.45×10^{-4}	50
YCR005C	160	1.68×10^{-6}	2.67×10^{-2}	1.23×10^{-4}	26
YCR011C	157	3.23×10^{-6}	1.84×10^{-2}	1.77×10^{-3}	54
YCR012W	157	3.23×10^{-6}	1.84×10^{-2}	1.77×10^{-3}	54
YCR014C	157	3.23×10^{-6}	1.84×10^{-2}	1.77×10^{-3}	54
YCR024C	112	2.90×10^{-3}	1.70×10^{-2}	5.15×10^{-4}	41
YCR027C	85	1.52×10^{-2}	4.20×10^{-3}	1.55×10^{-4}	35
YER129W	152	3.31×10^{-4}	9.76×10^{-3}	4.84×10^{-3}	23
YER133W	121	6.65×10^{-3}	1.28×10^{-3}	2.72×10^{-2}	15

Table 2.4 (continued)

pORF[1]	#Total[2]	prob C[3]	prob F[4]	prob P[5]	#matching[6]
YER138C	89	1.77×10^{-5}	2.75×10^{-5}	8.27×10^{-5}	8
YGR289C	7	4.01×10^{-2}	3.09×10^{-4}	4.20×10^{-4}	2
YHR005C-A	117	2.19×10^{-6}	5.27×10^{-4}	5.98×10^{-3}	37
YLR155C	17	5.62×10^{-8}	8.86×10^{-9}	7.65×10^{-9}	4
YLR157C	17	5.62×10^{-8}	8.86×10^{-9}	7.65×10^{-9}	4
YLR158C	17	5.62×10^{-8}	8.86×10^{-9}	7.65×10^{-9}	4
YLR160C	17	5.62×10^{-8}	8.86×10^{-9}	7.65×10^{-9}	4
YLR258W	201	3.69×10^{-4}	2.14×10^{-4}	4.56×10^{-3}	98
YLR260W	200	4.38×10^{-4}	2.08×10^{-4}	4.45×10^{-3}	97
YLR262C	206	1.62×10^{-3}	2.42×10^{-4}	5.02×10^{-3}	99
YLR450W	11	4.60×10^{-2}	8.46×10^{-3}	3.79×10^{-2}	2
YNL067W	298	4.72×10^{-19}	1.11×10^{-6}	4.29×10^{-4}	39
YNL069C	370	4.42×10^{-23}	9.52×10^{-8}	4.76×10^{-5}	47
YNL070W	370	6.91×10^{-31}	1.04×10^{-2}	4.76×10^{-5}	11
YNL071W	370	1.34×10^{-19}	7.21×10^{-3}	3.40×10^{-2}	121
YNL072W	370	1.34×10^{-19}	7.21×10^{-3}	3.40×10^{-2}	121
YNL073W	395	8.19×10^{-21}	2.46×10^{-2}	3.07×10^{-6}	127
YNL079C	562	1.86×10^{-37}	2.57×10^{-4}	7.00×10^{-14}	13
YNL081C	575	6.92×10^{-51}	1.92×10^{-17}	1.64×10^{-14}	81
YNL082W	575	1.44×10^{-3}	1.25×10^{-6}	3.63×10^{-14}	208
YNL088W	566	1.25×10^{-3}	9.48×10^{-8}	9.30×10^{-15}	204
YNL090W	536	7.99×10^{-4}	3.90×10^{-8}	9.43×10^{-15}	196
YNL093W	430	2.36×10^{-3}	1.92×10^{-7}	7.84×10^{-11}	164

Table 2.4 (continued)

pORF[1]	#Total[2]	prob C[3]	prob F[4]	prob P[5]	#matching[6]
YNL097C	327	6.29×10^{-3}	8.14×10^{-4}	4.87×10^{-8}	130
YNL098C	268	7.84×10^{-14}	1.22×10^{-4}	1.07×10^{-6}	109
YNL099C	268	7.84×10^{-14}	1.22×10^{-4}	1.07×10^{-6}	109
YOL086C	399	7.28×10^{-12}	3.14×10^{-4}	2.71×10^{-3}	23
YOR130C	101	5.10×10^{-12}	2.37×10^{-31}	9.48×10^{-10}	32
YOR136W	61	4.23×10^{-2}	8.79×10^{-8}	1.01×10^{-4}	10

[1] pORF with GO terms in all three ontologies matching over-represented GO terms detected in the vertical list above it.

[2] total number of eQTLs in the vertical list over the pORF used by TermFinder. Some eORFs with unknown annotation were dropped by TermFinder.

[3] corrected (for multiple tests within a vertical list above a pORF) P value reported by GO:termFinder package. This P value is associated with a GO term in Cellular Component ontology. The GO term annotates an eORF subgroup in a vertical list over the pORF. When multiple over-represented GO terms (and hence multiple P values) were found for a same subgroup of eORF, the minimal P value was reported.

[4] corrected P value for Molecular Function ontology.

[5] corrected P value for Biological Process ontology.

[6] number eORFs in the subgroup with matching over-represented GO terms.

The popularity of full matching cases within eQTL hot-spots (Fig 2.2) suggests that trans-acting transcriptional control in this mapping population can be executed by genes' 'coworkers', or 'indirect transcriptional effects' from 'perturbed pathways' (Brem *et al.*, 2002), as well as by their supervisors: transcriptional factors. For each of these 50 full matching cases, we suggest the sequence variation around the pORF (as appears in the first column of table 2.4) introduces functional perturbation to the local cellular network around the pORF. As the effect is disseminated along the pathways, the neighboring genes along the pathways will response to it in various ways, which include adjusting their own transcript abundance.

2.4 DISCUSSION

2.4.1 Mapping Strategy

LRT based interval mapping strategy still has a few advantages when compared with a direct correlation study between trait variation and marker variation, even in this mapping population with dense markers. The underlying EM algorithm can handle very well the situation of missing markers and occasional large marker intervals. It provides a straightforward way to find an interval estimate of eQTL location. The speed of calculation is still acceptable. It took about one day in a ten node cluster machine to get all LRS for eQTLs of 6195 traits in a data set with about 3000 markers and 112 individuals.

Permutation tests can deal with the multiple test issue in a genome scan for one trait, but not for all traits. Here we provide some intuitive explanation for why more than 10000 (summation of '#scanned' column in table 2.1) such tests would end up with FDR at 8%. For example, in the first cycle of MIM, with type I error rate 5%, the up bound of expected number of false eQTLs is roughly $6195 \times 5\% = 310$. Such a number could be devastating in certain cases, but not here. With 3354 declared eQTL (table 2.1), we can claim FDR is roughly $310/3354 = 9\%$. Similar argument can be applied to eQTL detected in other cycles. More rigorous computation could be found in the method section. The elevated type I errors due to multiple testing would be still acceptable in terms of FDR if we have sufficient positive discovery, which is the case for expression QTL mapping.

Storey *et al.* (2005) presented an elegant way to control FDR while performing model selection. However, Storey *et al.* (2005) declared only 170 traits with two eQTLs while controlling FDR at 10%. Different ways in pre-processing data and handling missing information do not seem sufficient to explain the discrepancy. A careful examination of their sequential search algorithm reveals: the best marker chosen at the second cycle of genome scan had either strong main effect or strong interaction effect with the first eQTL or both. We tried to replicate their procedure and found 182 traits with 2 eQTLs at FDR 10%, which shared 125 traits in common with Storey's 170 traits. It is noticed that the $25\%, 50\%, 75\%$ quartiles of F statistics for the second eQTL main effect (type I test) are $4.54, 9.94, 14.54$, respectively. For those second eQTLs with main effect F statistics less than 4.54, the mean F statistics for epistatic effect is 13.3. Thus, this search algorithm is targeted for interaction effects. However, if sacrificing this feature and modifying the original algorithm by considering only main effects when searching the second eQTL, which is fairly close to MIM procedure, we found a two QTL model with main effects only for 729 traits while controlling FDR at 10%. There was fewer number of strong interaction effects between these eQTL pairs (Figure 2.4), but apparently we traded them for more main effects.

Storey's method is able to control FDR for all the two eQTL models, which is statistically very attractive but also means that no inference is made when there is only one eQTL in the model. The algorithm will search for a second eQTL no matter how small the maximal test statistic for the first eQTL is, which would be excluded in MIM. Thus, by making decision at each step of sequential search, MIM effectively reduces the number of unnecessary tests: if a trait is unlikely to have its first eQTL, we just do not search for the second eQTL. The reduction would be more significant in later cycles of sequential search (table 2.1). In terms of Bayesian estimate of FDR (Efron *et al.*, 2001), it is to reduce the prior probability that a test statistic belongs to a false eQTL. In terms of exploring the solution space, it is to discard certain part of the space where the chance that the best solution is in it will be low. This is a feature shared by many heuristic search algorithms. We tried to altered his algorithm by only searching for the second eQTL when the F statistic of the first QTL is larger than 13.94. We picked this value just to show what we could gain if restricting the search for the second QTL. Controlling FDR at 10%, a two QTL model was declared for 704 traits. Because more traits were declared with a two QTL genetic model, and interaction effects were taken into account during genome scans, the number of strong interaction effects detected using the restricted search was even larger than that from the original method in Storey *et al.* (2005) (Figure 2.4).

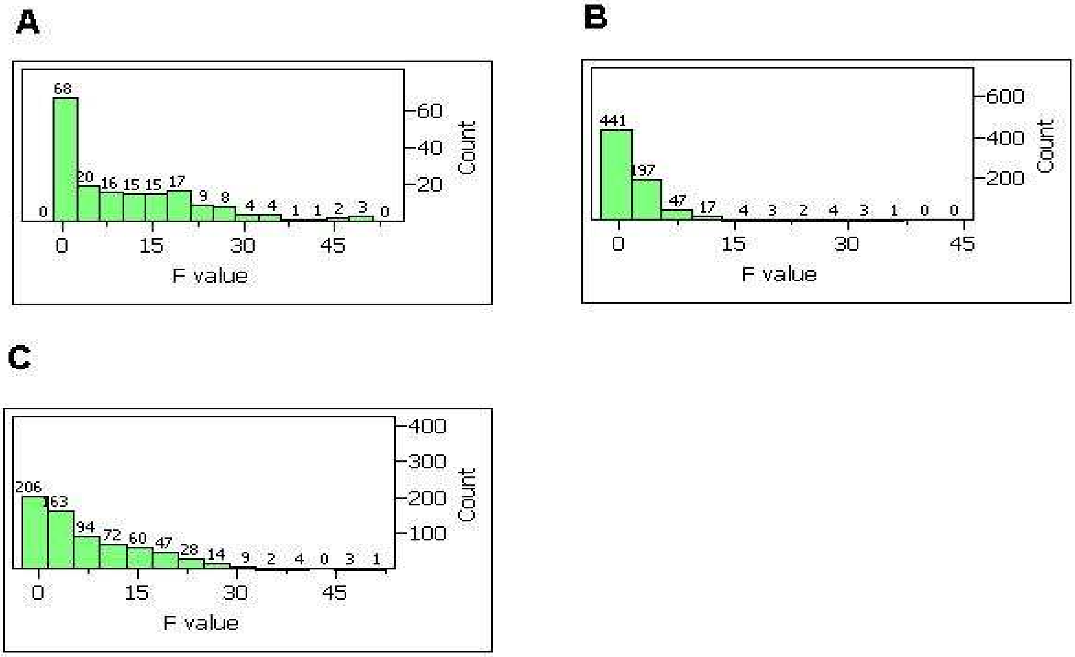

Figure 2.4: Histograms of F statistics of interaction effects between two QTLs detected through various methods. Numbers above each bar show the number of traits, or the number of two QTL interaction effects with type I F statistics in the bin. A: The significance of the interaction effects in the two QTL models claimed using the original method from Storey *et al.* (2005). B: Based on the two QTL models from sequential searches involving only main effects, the significance of the interaction between each QTL pair for a trait was tested. C: The significance of the interaction effects in the two QTL models, while restricting the search for the second QTL according to the F statistics of the first QTL.

2.4.2 QTL Mapping as a Genome Annotation Tool

In Figure 2.2, for each eORF, there are a horizontal list of genes in its eQTL, which are much more likely to include the causes of expression variation of the eORF compared with the genome background. In the meanwhile, above each pORF, there are a vertical list of genes, each of which has a transcript abundance profile co-segregating with the sequence variation around the pORF.

These horizontal lists were obtained through eQTL mapping's ability to classify genomic region according to its relevance to the expression variation. They can be easily incorporated into other bioinformatics studies such as a probabilistic view of gene relationship (Fraser and Marcotte, 2004). These vertical lists demonstrated the ability of eQTL mapping, in the context of genetical-genomics study, to identify co-regulated gene expression clusters. Central Dogma suggests that the genomic segments delineated by recombination events serve as pivotal random variables in studying the correlation structure of expression profiles.

2.4.3 Scale-free network

Many cellular networks, including both protein-protein interaction (Han *et al.*, 2004) and regulatory network by transcriptional factors (Luscombe *et al.*, 2004), belong to a scale-free network. (Barabasi and Oltvai, 2004). Let k be the number of links from one node (gene) to the rest of the genome; p be the probability distribution of k. The characteristic of a scale-free network is that most genes have very few connections, while a small number of genes are directly connected to a large number of genes; or numerically, k has a power-law distribution, i.e., $\Pr(K = k) \propto k^{-r}$, where r is a positive constant. On the contrary, $\Pr(K = k) \propto e^{-k}$ is for a random network. In our case, most of pORFs were covered by one or two eQTLs, while a few genomic segments affect as many as 600 expression traits. The probability density, p, for a pORF to have k eORFs mapped onto it could be well fitted with a function proportional to $k^{-1.66}$ (index was obtained by linear regression of $log(p)$ onto $log(k)$, see Figure 2.5). Compared with a random network, a scale free network has a much shorter mean path length to connect gene pairs, allowing a local fluctuation in biochemical pathways easily amplified to hundreds of its near or far neighbors. Thus, a single sequence variation can have pleiotropic effects on hundreds of genes' expression. Such a wide spread change in the transcriptome has certain potential to buffer the adverse effect associated with the sequence variation. In this way, a high heritable expression variation can be maintained in the population because of its minor effect on fitness (Barton and Keightley, 2002).

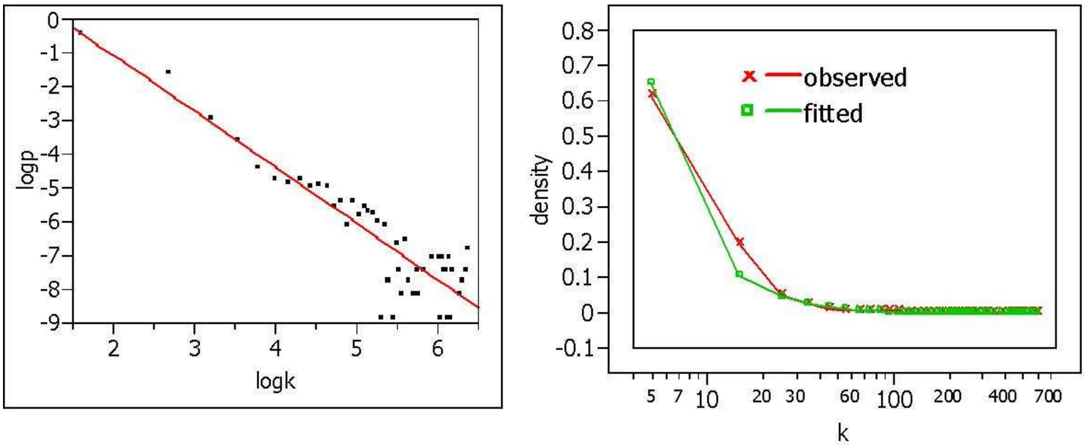

Figure 2.5: QTL mapping results suggest transcriptional regulation network follows scale free property. After collecting the numbers of expression traits mapped onto each pORF, these numbers were put into a frequency table with bin size of 10. Let p be the frequency and k be the mid-bin value. Based on the frequency table, we tried to estimate the empirical density function using parametric curve fitting. Left: Log–transformed frequencies $log(p)$ were regressed onto log-transformed mid-bin values $log(k)$. The linear regression function is $logp = 2.24 - 1.66 * logk$. Right: Observed and fitted density function: $\Pr(k) = e^{2.244} \times k^{-1.66}$.

2.5 References

Ashburner, M., C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, 2000 Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

Balakrishnan, R., K. Christie, M. Costanzo, K. Dolinski, S. Dwight, S. Engel, D. Fisk, J. Hirschman, E. Hong, R. Nash, R. Oughtred, M. Skrzypek, C. Theesfeld, G. Binkley, C. Lane, M. Schroeder, A. Sethuraman, S. Dong, S. Weng, S. Miyasato, R. Andrada,

D. Botstein, , and J. Cherry, 2005 Saccharomyces Genome Database. Ftp://ftp.yeastgenome.org/yeast/.

Barabasi, A.-L., and Z. N. Oltvai, 2004 NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION. *Nat Rev Genet* **5**: 101–113.

Barton, N. H., and P. D. Keightley, 2002 UNDERSTANDING QUANTITATIVE GENETIC VARIATION. *Nat Rev Genet* **3**: 11–21.

Basten, C., B. Weir, and Z. Zeng, 2002 *QTL Cartographer, Version 1.17*. Department of Statistics, North Carolina State University, Raleigh, NC.

Bing, N., and I. Hoeschele, 2005 Genetical Genomics Analysis of a Yeast Segregant Population for Transcription Network Inference. *Genetics* : genetics.105.041103.

Brem, R., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

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. *Nat Genet* **37**: 233–242.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285–94.

Efron, B., and R. Tibshirani, 2002 Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol* **23**: 70–86.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher, 2001 Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**: 1151–1160.

Fraser, A. G., and E. M. Marcotte, 2004 A probabilistic view of gene function. *Nat Genet* **36**: 559–564.

Han, J.-D. J., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, 2004 Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**: 88–93.

Ihmels, J., R. Levy, and N. Barkai, 2004 Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat Biotech* **22**: 86–92.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, 2004 The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–80.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* **152**: 1203–1216.

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

Lee, I., S. V. Date, A. T. Adai, and E. M. Marcotte, 2004 A Probabilistic Functional Network of Yeast Genes. *Science* **306**: 1555–1558.

Luscombe, N. M., M. Madan Babu, H. Yu, M. e. Snyder, S. A. Teichmann, and M. Gerstein, 2004 Genomic analysis of regulatory network dynamics reveals large topological changes. *nature* **431**: 308–312.

Marchini, J., P. Donnelly, and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**: 413–417.

Mewes, H., D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil, 2002 MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**: 31–34.

Sax, K., 1923 THE ASSOCIATION OF SIZE DIFFERENCES WITH SEED-COAT PATTERN AND PIGMENTATION IN PHASEOLUS VULGARIS. *Genetics* **8**: 552–560.

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.

Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440–5.

Wolfinger, R., G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**: 625–637.

Yvert, G., R. B. 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 rol e of transcription factors. *Nat Genet* **35**: 57–64.

Zeng, Z., C. Kao, and C. Basten, 1999 Estimating the genetic architecture of quantitative traits. *Genet Res.* **74**: 279–289.

Chapter 3

eQTL Viewer: visualizing how sequence variation affects transcription

3.1 Abstract

3.1.1 Background

Expression Quantitative Trait Locus analysis methods have been used to identify the genetic basis of variation in gene expression. In such studies, thousands of expression profiles are related with marker data from thousands of sequence regions and each sequence region can include many genes. There is a need for new tools to explore these results in a way that looks across many genes at once.

3.1.2 Results

We have developed a web-based tool to visualize the relationships inferred by such an analysis in Scalable Vector Graphics. The resulting plot is able to display genomic data with high resolution, and dynamically superimpose biological annotation provided by users onto mapping results.

3.1.3 Conclusions

Our tool provides an efficient and intuitive way for bioinformaticians to explore expression Quantitative Trait Locus results, and to ask questions about the biological mechanisms and transcriptional regulation patterns suggested by the analysis.

3.2 Background

Transcriptional control is a crucial step in development and cellular environmental response. Recent studies have demonstrated that mRNA levels vary in both natural and experimental populations (Oleksiak *et al.*, 2002). Expression Quantitative Trait Locus (eQTL) mapping seeks to explain this variation by identifying relationships between transcript abundance and specific genomic markers (Schadt *et al.*, 2003). While classical QTL mapping focuses on one or few traits, an eQTL mapping study may have thousands of expression traits. Since several marker regions may affect each expression trait, many thousands of eQTLs are often identified. The patterns among these eQTLs allow us to ask questions that we cannot address by the traditional approach. For example, this type of analysis has been used to gain insights into the relative proportion of cis- and trans- acting regulatory regions in yeast (Brem *et al.*, 2002; Brem and Kruglyak, 2005; Yvert *et al.*, 2003), mice and rat (Bystrykh *et al.*, 2005; Chesler *et al.*, 2005; Hubner *et al.*, 2005). Such findings provide a starting point for additional exploration. A marker region that associates with several expression traits for genes belonging to a specific pathway may suggest a master transcriptional regulation region for that pathway. A sensible next step is to query a genomic database to find if there is a candidate transcription factor located in that region. Inversely, what does it mean if many distinct sequence regions are linked to the expression of a particular gene? These analyses require the ability to view eQTL results at a variety of scales: across all expression traits, among the genes of a single pathway, and at the level of a single gene. Additionally, researchers need a quick and straightforward means to explore genes that comprise interesting features. We have addressed these needs with the software tool we present here. We have developed eQTL Viewer, a web-based tool that presents eQTL mapping results visually. We display the results for thousands of expression traits in a single plot, which makes features such as trans- and cis- regulatory readily identifiable. We extend the basic plot with the ability to link graphical features and external bioinformatics resources, thus providing an intuitive platform for discovery. See figure 3.1 for the role of eQTL Viewer in eQTL mapping.

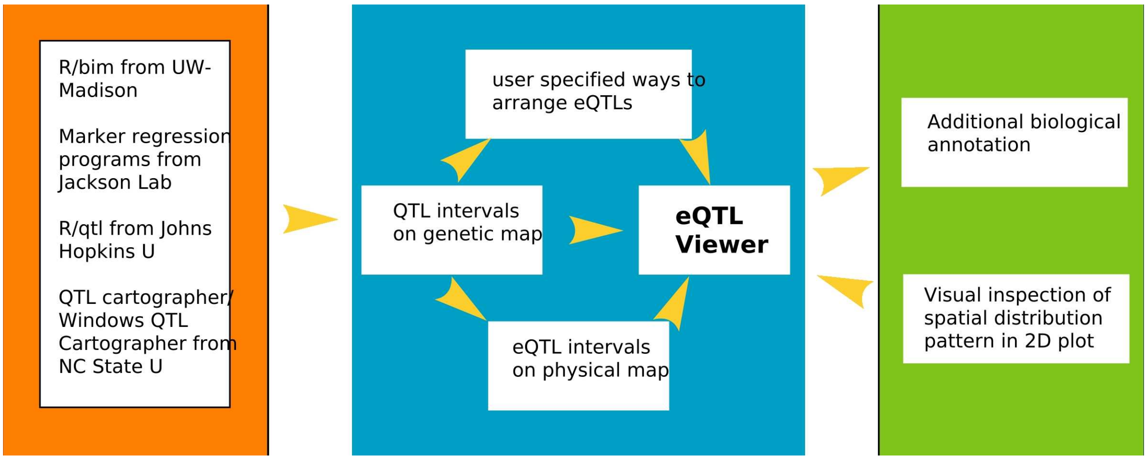
R/bim from UW-Madison
Marker regression programs from Jackson Lab
R/qtl from Johns Hopkins U
QTL cartographer/ Windows QTL Cartographer from NC State U
user specified ways to arrange eQTLs
QTL intervals on genetic map
eQTL Viewer
eQTL intervals on physical map
Additional biological annotation
Visual inspection of spatial distribution pattern in 2D plot

Figure 3.1: Role of eQTL Viewer in eQTL mapping

3.3 Implementation

eQTL Viewer leverages the power of Scalable Vector Graphics (SVG), an open standard for graphical display recommended by World Wide Web Consortium (W3C) `http://www.w3.org/TR/SVG/`. Instead of creating a static graphic, SVG is a set of instructions for drawing a graphic written in eXtensible Markup Language (XML). As such, an SVG graphic element can be viewed at a wide variety of resolutions without sacrificing quality. Virtually unlimited text information can be associated with each graphic element and users can control what information is displayed at one time. These features make SVG well suited for visualizing genomic data, and SVG has been used for a variety of bioinformatics applications (Tanoue *et al.*, 2002).

eQTL analyses produce a list of contiguous genetic markers significantly associated with the expression trait of each gene. Alternately, users can identify genes corresponding to these markers and describe an eQTL as a list of genes. We offer auxiliary programs to do such a transformation in our website. These data are input into eQTL Viewer in a simple XML format that is detailed online. This basic data are converted into a graph with all expression traits on the vertical axis and genetic markers (or genes) on the horizontal axis. This displays the relationships between marker regions and expression traits inferred by eQTL mapping and allows biologists to explore them at the whole genome level.

Several novel features extend the usefulness of the graph. Users can supply known gene-gene relationships and highlight all correspondence between them and eQTL mapping results. Users can search and highlight QTLs in the graph by their trait gene names. One use of this function is to highlight all QTLs associated with a metabolic pathway. Users can also rearrange expression traits along the vertical axis as their research hypotheses. For example, eQTLs associated with a specific pathway can be clustered together. Each genetic element in the graph can be linked to annotation in an external database such as NCBI. This provides a framework for exploratory analysis in which the user can drill down on genes of interest and use the linked biological information to interpret results. New and customizable annotation easily can be superimposed onto the graph. The strength of this approach is that the graph becomes a platform for synthesis as additional biological knowledge becomes available. All these features make eQTL Viewer a unique tool for asking biological questions from eQTL mapping results.

3.4 Results

We applied multiple interval mapping (Kao *et al.*, 1999) to the data from Brem and Kruglyak (2005) to illustrate these features. We associated 5182 eQTLs with 3367 expression traits and used eQTL Viewer to display the result (Figure 3.2). eQTLs on the diagonal line are located at the same genomic location as an expression trait gene. This can be interpreted as a cis-acting regulatory region. Large number of trans-acting eQTLs cluster in certain regions of chromosome 2, 14 and 15 (Figure 3.2). As directly implied by eQTL analysis, transcript abundances of those genes with trans-acting eQTLs in a column may be associated with sequence variation of the same set of genetic factors, or different genetic factors in close linkage. Thus, it is not surprising to find that transcript abundance levels themselves are highly correlated. These two types of association suggest that the genomic region is biologically responsible for the correlated transcriptional pattern for the cluster of genes.

Using features described above, the graph shows 49 cases where a trait gene's eQTL contains the transcriptional factor for the gene (Luscombe *et al.*, 2004). Such a matching case suggests a hypothesis that the genomic region encoding the transcriptional factor contains causal sequence variations for the expression variation of the trait gene in this mapping population. eQTLs of genes in the Oxidative phosphorylation pathway (Kanehisa *et al.*, 2004) cluster in the middle of chromosome 15. The pathway traps the energy from the mitochondrial electron transport assembly to synthesize ATP, which provides extensive energy support for cell functions. Such a clustering pattern suggests a important regulatory control gene or genes affecting the pathway could be located there. These patterns can be found in our supporting website.

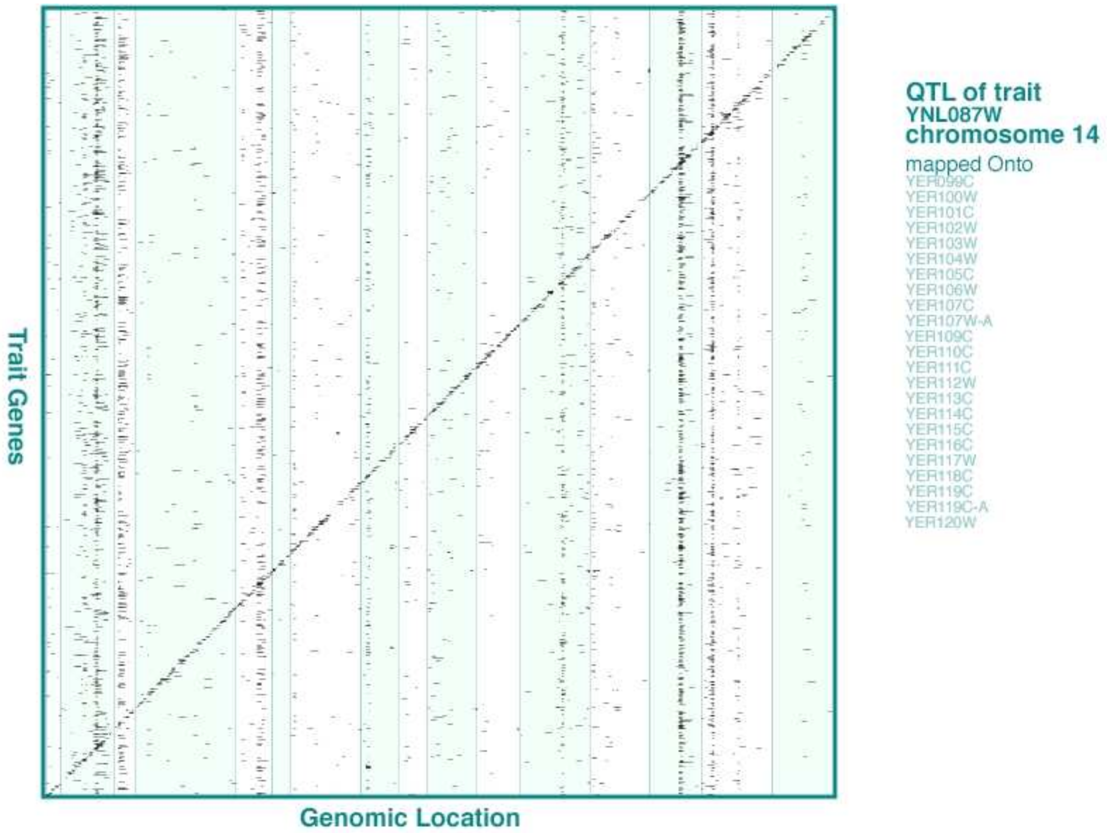

Figure 3.2: The graph shows all eQTLs as small bars. Each eQTL's associated trait gene determines its vertical coordinate, while genes in the eQTL determine its horizontal coordinate. These genetic features are arranged along both axes according to their physical location. The user can zoom in on regions of interest. When the mouse pointer is over a specific eQTL, the names of genes in the eQTL, and the name of the expression trait gene appear in the right sidebar.

3.5 Discussion

We emphasize gene-gene relationships in eQTL Viewer, which are the key questions biologists will ask after eQTL mapping. eQTLs can span physically large genomic regions, and the precision of eQTL location is limited by the experimental design. Due to these considerations it is difficult to know exactly which gene within an eQTL is the gene of interest (Mackay, 2001). This presents a dilemma to the eQTL community, because the many public bioinformatics resources are gene-centric. By linking expression traits and physical markers to their corresponding genes, our implementation will help scientists conceptualize each eQTL as a list of pairwise relationships between a trait gene and the multiple genes in the eQTL. This goes a step further than just showing the relationship between RNA probes and polymorphic markers. Combining experimental results with external sources is the essence of exploration in the age of bioinformatics.

Mueller *et al.* (2006) recently introduced their eQTL Explorer package in a similar spirit. While both software packages provide features for exploring this eQTL results, they have differing approaches and fulfil complementary functions. While eQTL Explorer excels as a clearinghouse and management tool for eQTL data, eQTL Viewer uniquely addresses the problem of visualizing eQTLs for the complete transcriptome simultaneously. Capturing thousands of traits and genome-wide markers in one plot emphasizes the complex interactions in transcriptional control as graphic features. With eQTL Viewer one can see clearly not only where a particular trans-acting eQTL is located in the genome, but also every expression trait associated with that eQTL.

3.6 Conclusions

eQTL Viewer is a simple and robust web service that generates a scalable graph to visualize such relationships between genotype and expression profile. It is our intent to help form a bridge between quantitative genetic analysis and systems biology by superimposing biological information onto eQTL mapping results.

3.7 References

Brem, R., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

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. *Nat Genet* **37**: 233–242.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz, F. Maciver, M. Mueller, O. Hummel, J. Monti, V. Zidek, A. Musilova, V. Kren, H. Causton, L. Game, G. Born, S. Schmidt, A. Muller, S. A. Cook, T. W. Kurtz, J. Whittaker, M. Pravenec, and T. J. Aitman, 2005 Integrated transcriptional profiling and linkage analysis for identificat ion of genes underlying disease. *Nat Genet* **37**: 243–253.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, 2004 The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–80.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* **152**: 1203–1216.

Luscombe, N. M., M. Madan Babu, H. Yu, M. e. Snyder, S. A. Teichmann, and M. Gerstein, 2004 Genomic analysis of regulatory network dynamics reveals large topological changes. *nature* **431**: 308–312.

Mackay, T. F., 2001 Quantitative trait loci in Drosophila. *Nat Rev Genet* **2**: 11–20.

Mueller, M., A. Goel, M. Thimma, N. J. Dickens, T. J. Aitman, and J. Mangion, 2006 eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. *Bioinformatics* **22**: 509–11.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford, 2002 Variation in gene expression within and among natural populations. *Nat Genet* **32**: 261–6.

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.

Tanoue, J., M. Yoshikawa, and S. Uemura, 2002 The GeneAround GO viewer. *Bioinformatics* **18**: 1705–6.

Yvert, G., R. B. 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 rol e of transcription factors. *Nat Genet* **35**: 57–64.

Chapter 4

A Bayesian Approach to Superimpose Biological Annotation to eQTL Mapping Results

4.1 INTRODUCTION

Transcriptional control is one of the most important steps for an organism to express the genetic information stored in its sequence as well as to respond to environmental changes (Ihmels *et al.*, 2004). Recent advance of genomic technology has made it possible to quantify transcript abundance systematically, as well as to genotype genetic markers covering the whole genome in a segregating population. Provided with these tools, expression Quantitative Trait Locus (eQTL) analysis has been applied in yeast (Brem *et al.*, 2002; Yvert *et al.*, 2003), drosophila (Schadt *et al.*, 2003) and mouse (Bystrykh *et al.*, 2005; Chesler *et al.*, 2005) to find chromosomal regions (i.e. eQTL) affecting the expression variation. Statistical procedure involved in eQTL mapping would guarantees that eQTL are significantly more likely to harbor biological elements affecting the expression traits than the rest of the genome. However, huge amount of experimental work will be needed before we can know which gene in an eQTL actually underlies the biological processes that cause the phenotypic variation in the mapping population (Mackay, 2001).

Biochemistry and genetics studies have shown that gene expression is regulated by cis-acting elements (such as in promoter regions) and trans-acting factors. Gene products of trans-acting factors, such as a transcriptional factor, may work directly on the cis-acting elements of their regulation targets. Also, mechanisms like feedback control allow indirect 'trans-acting transcriptional effects' in 'perturbed pathways' (Brem *et al.*, 2002).

The advantage of performing eQTL mapping in well studied organisms is that knowledge through previous biological investigation can help eQTL experimenters to have certain biological interpretation of their mapping results before further experimental investigation. For these organisms, an eQTL can readily be transformed into a one (trait) gene to many genes (in the trait genes' eQTL) relations, or a list of pairwise gene relationships. When a gene pair from such a list is found to be involved in same biological processes or functionally related according to certain annotations, an annotation can be super-imposed onto the eQTL. The annotation could serve as a plausible biological explanation for the QTL and a hypothesis to test in future experiments.

For example, trait gene's genomic location is a rudimental annotation, which has been used by most published eQTL analysis to find cis-acting eQTL. Most biological network topology can be reduced to a set of related gene pairs. In yeast, comparison between known transcriptional regulatory network composed of transcriptional factor-target pairs and eQTL showed few matching cases (Yvert *et al.*, 2003), which made indirect transcriptional effect an important alternative mechanism.

There are several disadvantages in these approaches.

- Superimposition based on string comparison of annotations is 'hard' in the sense that it will only give result of 'yes' or 'no'. Thus, multiple superimpositions for an eQTL are treated similarly as 'yes'.
- Statistical patterns are treated equally. Lists of pairwise gene relationships from different eQTLs are compared with biological annotations in the same way. However, each eQTL is declared with different likelihood ratio statistics.
- Pairwise biological annotations are treated equally. Neighborhood connectivity is barely taken into account when the known relationship between

a gene pair is compared with statistical patterns. Since the result of comparison is either ‘yes’ or ‘no’, delicate network structures in annotation are simplified as ‘related’ or ‘not related’ pairwise gene relationships.

- When several sources of annotations can be superimposed onto an eQTL, there is no ready way to weight different types of annotations differently. Annotations are created using different technologies, aiming to detect different aspects of biological activities. They might overlap with each other, but have more or less unique information of their own.

In this paper, using the yeast data published by Brem and Kruglyak (2005) as an example, we present a Bayesian approach to prioritize a list of genes in an eQTL with loose or strong functional relationship with the trait gene. In our Bayesian analysis framework, statistical significance associated with an eQTL provides information to estimate the prior probability that a gene is a biological factor underlying the statistically detected eQTL; global biological connection scores (GCS) between a gene and the corresponding trait gene can be used to update the prior to give the posterior probability. We inferred GCS from pairwise scores for functional relationship among yeast gene pairs in finalNet (Lee *et al.*, 2004), which incorporates annotation information from different biological experiments and databases.

4.2 METHODS

The yeast data (Brem and Kruglyak, 2005) include 6195 unique expression traits assayed in 112 haploid yeast obtained by crossing a standard laboratory strain and a wild strain isolated from a California vineyard. 2956 SNPs markers were genotyped across the genome.

4.2.1 eQTL Mapping

We applied Multiple Interval Mapping (Kao *et al.*, 1999) (MIM) to search the yeast genome for the best statistical model for each expression trait. Multiple eQTL could be included in a model through a series of genome scans. In each cycle of genome scan for an expression trait, likelihood ratio tests were performed on and between markers conditional upon QTL found in previous cycles. No scan would be performed for a trait if no eQTL was found for the trait in the previous cycle of genome scan. The maximal likelihood ratio statistic (LRS) in one pass of genome scan will give rise to an eQTL if it is above certain threshold. We used the 95% quantile of genome-wide null LRS distribution as the threshold, to declare the 1.5 LOD support interval around a LRS peak as an eQTL. SNPs with known position in both the genetic and physical maps serve as anchors for eQTL regions in translating eQTL genetic map intervals into chromosome regions with physical boundaries. Gene lists located in the regions were then produced.

4.2.2 Prior Probability of not Being an Underlying Gene

To adapt current 'hard' superimposition method to a Bayesian approach, we summarized eQTL mapping results as prior probabilities. We indexed genes in the genome as g_i. Since eQTL are chromosome segments, g_i usually means a gene in an eQTL. We indexed traits as e_k. Since in eQTL study, each trait is the mRNA level of a gene, e_k also refers to the k-th gene whose expression level is a trait. Thus, g_i and e_k are sampled from the same gene pool. For each e_k, denote its j-th eQTL region as $R_{k,j}$, which contains $\gamma_{k,j} > 0$ candidate genes which may or may not be the underlying gene affecting e_k but are all physically located in the region $R_{k,j}$. Define an index variable Q_{kj} associated with $R_{k,j}$, where $Q_{kj} = 1$ denotes that $R_{k,j}$ contains at least one underlying gene, or that $R_{k,j}$ declared through the mapping procedure is a 'true' eQTL; and $Q_{kj} = 0$ for otherwise. $\Pr(Q_{kj} = 1)$ is related to a parameter p_{kji}, the prior probability that a particular gene g_i in eQTL $R_{k,j}$ is not the underlying gene, as following:

$$\Pr(Q_{kj} = 1) = 1 - \prod_{i=1}^{\gamma_{kj}} p_{kji}, \tag{4.1}$$

where we assume genes affect e_k independently. p_{kji} represents eQTL mapping results as prior information to our Bayesian analysis. eQTL analysis has little power to discriminate genes within an eQTL in terms of their relevance to the trait: variation of LOD scores at different position within an eQTL is less than 1.5 according to our mapping procedure. Thus, we assume p_{kji} is identical for different g_i in $R_{k,j}$ and denote the prior probability as $p_0^{(kj)}$, or p_0 to simplify the notation while keeping in mind that p_0 differs among eQTL.

Storey *et al.* (2005) described an empirical Bayesian method to find the posterior probability of an eQTL to be true given its test statistic in the framework of multiple locus linkage analysis. Their method attaches Bayesian probabilities for both individual eQTL in a multiple eQTL model and the overall model for each expression trait. We applied their method on our MIM results. For a list of maximal *LRS* obtained in one cycle of genome scan, the prior probability that a *LRS* in the list is associated with a false eQTL was calculated according to Efron *et al.* (2001). More specifically, all *LRS* that are smaller than the lower quartile of the list were considered as purely generated by false eQTL. The estimated prior probabilities range from 0.33 to

0.5, each of which is larger than the corresponding estimate using QVALUE package (Storey and Tibshirani, 2003), resulting in a conservative estimation of posterior probabilities. We did not use the estimation from QVALUE because it is sometimes unlikely small. Then, $\Pr(Q_{kj} = 1|LRS)$, the posterior probability that $R_{k,j}$ is true given its LRS was obtained using non-parametric logistic regression method. This procedure is a variant of the empirical Bayesian method introduced by Efron *et al.* (2001) and was first applied to QTL analysis by Storey *et al.* (2005). See section 1.6.4 for computational details. We used this posterior probability to estimate $\Pr(Q_{kj} = 1)$, and hence p_0 using formula (4.1).

4.2.3 Connection Scores Between Gene Pairs

A connection score between a gene pair quantifies their functional relatedness suggested by biological annotations. Transforming annotations to scores makes them amenable for numerical processing and probabilistic superimposition.

Ideally, a connection score would be a summary statistic of comprehensive biological observations involving a gene pair. A possible way to obtain such scores is to use semantic similarity measure of GO terms associated with gene pairs (Lord *et al.*, 2003). However, finalNet (Lee *et al.*, 2004), an available weighed connected graph, seems a better choice to illustrate our algorithms. It nicely weights and combines heterogeneous knowledge (co-expression, protein interaction inferred from sequence similarity and biochemical/genetical experiments) from yeast studies. The graph contains 5545 yeast genes as nodes. A log likelihood score (LLS) between a gene pair forms a weighted edge between the nodes. LLS quantifies 'functional linkage' between a gene pair supported or not supported by various evidences. A study of their algorithm suggested that the computation of a LLS only takes into account the evidences regarding to the gene pair under question. A similar argument can be applied to scores from semantic similarity based upon GO terms.

We call scores like LLS 'direct connection scores' to emphasize that they represent pairwise functional relatedness without considering the global network structure. In the following text, we will introduce globalized connection scores, or GCS, denoted as $C(e_k, g_i)$, between a trait gene e_k and $g_i, g_i \in R_{k,j}$. GCS will be larger if there are more gene nodes clustered around the gene pair (e_k, g_i) with edges connecting each other. The following biological considerations suggest that GCS are more appropriate in our application.

- The complex topology of biological networks (Ihmels *et al.*, 2004; Luscombe *et al.*, 2004) indicates that a gene's transcriptional activity is modulated globally: all active regulators in the genome will act simultaneously.
- Mechanisms like hierarchical regulation (Barabasi and Oltvai, 2004) and feedback control suggest that regulation can be exercised directly between a gene pair, or indirectly through a chain of genes on a pathway.

To convert direct connection scores into GCS, we made slight modification of RANKPROP algorithm by Weston *et al.* (2004). Inspired by google's PAGERNK algorithm to infer global internet structure from hyperlinks among individual web pages, Weston *et al.* (2004) presented RANKPROP algorithm to globally rank sequence similarity between a query protein and the rest of proteins, which had been already connected with weighted edges of similarity scores from BLAST or PSI-BLAST. Treating finalNet as a similar weighted network input, we applied the following algorithm.

Algorithm 1 Derive GCS from LLS

```
for all $e_k$ in genome do
    $\mathbf{Y}_k = \{y_i\}$, where
    Let $\mathbf{Y}_k^t = \{y_i^t\}$ be the value of $\mathbf{Y}_k$ in iteration $t$, $\mathbf{Y}_k^0 = 0$
    repeat
        for all $g_i$ in genome do
            if $i \neq k$ then
                $y_i^{t+1} = LLS(g_i, e_k) + \alpha \sum\limits_{j \neq i,k} LLS(g_i, g_j) \times y_j^t$
            else
                $y_i^{t+1} = 0$
            end if
        end for
        normalization: $y_i^{t+1} = \frac{y_i^{t+1}}{\sum\limits_i y_i^{t+1}}$
    until $\|\mathbf{Y}_k^{t+1} - \mathbf{Y}_k^t\| < 10^{-6}$ or $t > maxT$
    report $y_i$ 's rank in $\mathbf{Y}_k$ as $C(e_k, g_i)$
end for
```

In the algorithm, for each trait gene e_k, we tried to obtain a vector $\mathbf{Y}_k$. The i-th element of the vector, y_i, quantifies the globalized functional connection between g_i and e_k. In line 7 of the algorithm, an estimated value of y_i in cycle $t+1$ is updated by the summation of the direction connection ($LLS(e_k, g_i)$) and indirect connection between e_k and g_i through g_j. g_j can be any gene in the genome other than e_k and g_i. For a specific junction g_j, the indirect connection part is formed by the product of direction connection between g_i and g_j ($LLS(g_i, g_j)$) and globalized connection between g_j and e_k (y_j^t) obtained in previous cycle. α, a tuning parameter affecting the speed that indirect connection scores diffuse into global scores suggested by Weston *et al.* (2004), was set as 0.9. To ensure convergence, $\mathbf{Y}_k$ is normalized into a unit vector in line 12 after each cycle. Such normalization processes make connection scores in different $\mathbf{Y}_k$ not directly comparable. To solve the problem, rank statistics from $\mathbf{Y}_k$ are reported as GCS, which rank the global functional relatedness between a specific e_k and the rest of genes in the genome. As the result, GCS between two different genes range from 1 to 5544, with larger scores for stronger biological connections between gene pairs. A gene's connection score with itself was manually assigned as 5545. Maximal number of iteration, $maxT$, was set as 20. Most $\mathbf{Y}_k$s converged before 10 cycles. $LLS(e_k, g_i)$ was set as zero when direct connection between e_k and g_i is missing from finalNet.

4.2.4 Mixed and Null Distribution of Connection Scores

Transcriptional regulation is an important component of functional regulation. Genes participating in different biological processes are less likely to affect the expression of each other than genes in the same process. We will allow $C(e_k, g_i)$ to modify our belief about g_i as an underlying gene for expression variation of e_k in the mapping population, assuming the connection between functional relatedness and transcriptional regulation.

Assume a distribution f_1, composed of $C(e_k, g_i)$ where g_i affects e_k; and a distribution f_0 for GCS where g_i does not affect e_k. If functional relatedness between a gene pair is independent of whether there is transcriptional regulation between the pair, we expect to find f_0 is similar with f_1, or any mixture of f_0 and f_1. On the other hand, if closer functional relationship does suggest more possible transcriptional interference, the probability mass of f_1, or any mixture of f_0 and f_1, is expected to shift towards larger values of GCS compared with f_0. The difference between f_0 and f_1 is the ultimate reason that from the value of the GCS, we can extract useful information about whether a GCS is sampled from f_0 or f_1, and hence whether a gene affects its trait gene.

Since there could be one or more or zero gene in an eQTL that truly affects e_k, GCS between e_k and g_i in $R_{k,j}$ are sampled from a mixed distribution. Let $z_i \equiv C(g_k, g_i)$ for g_i in $R_{k,j}$:

$$f_m(z_i) = p_0 f_0(z_i) + (1 - p_0) f_1(z_i) \tag{4.2}$$

where p_0 is the probability that g_i is not an underlying gene as defined previously. f_m will vary as the change of p_0, or vary from eQTL to eQTL.

Just like we have declared high LRS regions from MIM as eQTL, low LRS regions for an expression trait are considered unlikely to contain genetic factors for the trait. For each g_i in $R_{k,j}$, we sampled 20 consecutive genes from low LRS regions of e_k. Their GCS with e_k are assumed to follow the distribution of f_0.

4.2.5 Posterior Probability Given Connection Score

With the previous definitions about f_0, f_1 and p_0, for expression trait e_k and g_i in $R_{k,j}$, we are able to assess the posterior probability whether an observed z_i is from f_1 rather than f_0 given z_i:

$$\frac{p_1 f_1(z_i)}{p_0 f_0(z_i) + p_1 f_1(z_i)} = 1 - p_0 \frac{f_0(z_i)}{f_m(z_i)}. \tag{4.3}$$

We arrange eQTL into 12 groups according to the p_0 associated with a gene in the eQTL, so that p_0 for different eQTL are similar within a group. Assuming f_m, as a function of p_0, does not change much as a small change of p_0 within each of the 12 groups, we applied non-parametric logistic regression method described by Efron *et al.* (2001) and Storey *et al.* (2005) to estimate f_0/f_m. Then, with formula (4.3), we can estimate the probability that z_i is from f_1, or the probability that g_i is the biological factor affecting e_k. It is actually the same method we used to estimate $\Pr\{Q_{kj} = 1\}$.

4.3 RESULTS and DISCUSSION

We only considered those genes included in finalNet. 4772 eQTL for 3109 expression traits were detected by MIM. *LRS* for eQTL vary from 15.4 to over 300. The number of genes in each eQTL varies from one gene to 123 genes, with an average at about 23. The whole mapping results were translated into 109668 pairwise relationships between trait genes and genes in the corresponding eQTL.

Figure 4.1A shows the histogram of p_0 associated with each of the 109668 pairs.

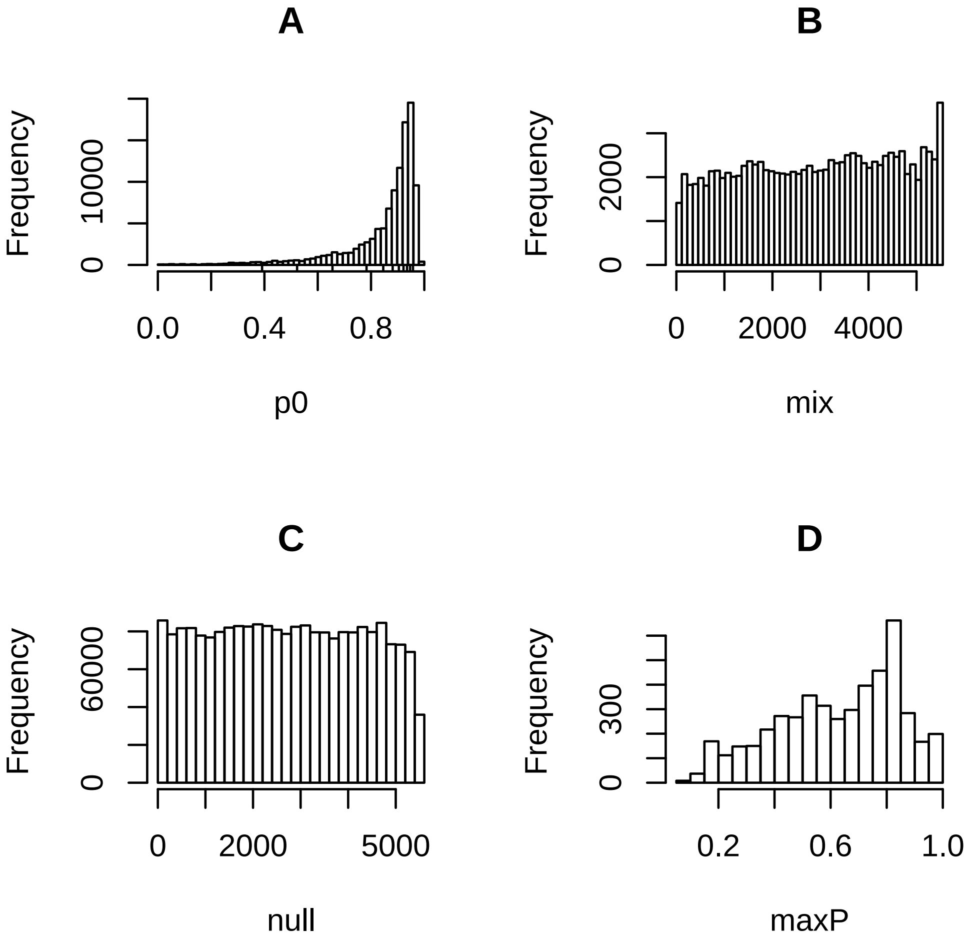

Figure 4.1: Results of probabilistic superimposition of annotations. A: histogram of p_0 for each gene in an eQTL not as an underlying gene. 11 small ticks between the graph and x-axis indicate how eQTL were groups into 12 groups B: pooled f_m. Individual f_m for each of the 12 groups varies, but not greatly from the pooled distribution. Figure C: pooled f_0. Figure D: histogram of maximal posterior probability being an underlying genetic factor among genes in an eQTL

p_0 was obtained from $\Pr(Q_{kj} = 1)$, which was estimated by $\Pr(Q_{kj} = 1|LRS)$. This probability turns out to be an increasing function of LRS. Together with formula (4.1), our procedure ensures that genes in eQTL with higher LRS and including less genes will have lower p_0. It is a very desirable feature. LRS is the ratio of the likelihood that there is an eQTL in the genomic region versus the likelihood of no eQTL there. The higher LRS, the more support from the data for the existence of an eQTL. Assuming number of yeast genes in a chromosome segment is approximately proportional to the length of the segment, eQTL including less genes tend to be shorter. An eQTL in the genetic map could be roughly interpreted as a confidence interval (CI) of the position of the genetic factor underlying the eQTL. Shorter CI generally means smaller sampling variability in estimating eQTL location. Thus, both higher LRS and shorter length deserve more credibility, and correspondingly, low p_0 for every gene in the eQTL.

As noted by Storey *et al.* (2005), $\Pr(Q_{kj} = 1|LRS)$ is conditional on assuming $Q_{kj'} = 1$ for all $j' \in [1, j-1], j > 1$. There could be some issues about using such a conditional probability to estimate $\Pr(Q_{kj} = 1)$. It is true that $\Pr(Q_{kj} = 1|LRS)$ depends not only on the significance of the current eQTL, but also on assumptions about the rest of eQTL declared for the same trait. However, LRS in a statistical model seems always associated with certain assumptions about other factors. In the case where there is only one eQTL found for a trait, we can still consider the LRS for this eQTL was obtained while assuming there were no other eQTLs for the trait. On the other hand, there is an apparent trend for multiple eQTL of the same trait declared through MIM to have little correlated genotypes, which results in a nearly balanced design, and orthogonality among LRS for multiple factors in a linear model.

We did not assess the probability of affecting a trait for genes outside the trait's eQTL. Those genes have essentially been assigned with negligible prior probabilities of being underlying genes, or $p_0 \rightarrow 1$, and then overlooked in Bayesian analysis. In this way, p_0 helps to preserve eQTL distribution pattern for the over 6000 expression traits. Most eQTL studies have showed that eQTL are distributed highly unevenly across the genome: most genomic regions contain one or two eQTL; a small portion are eQTL 'hot-spots', each affecting hundreds of traits. In our approach, genes in the former regions are assigned with $p_0 < 1$ for only a few traits; while genes in hot-spots will have intermediate p_0 for many traits.

Figure 4.1B shows f_m for GCS associated with 109668 pairwise relationships from eQTL study. The probability density has a little trend to increase as GCS increases. There is a sharp contrast when it is compared with f_0 (Fig 4.1C), whose density is flat for most scores, but drops suddenly when scores are larger than 5000. To estimate the posterior probability as formula (4.3), actually f_m and f_0 were estimated separately in each of the 12 groups. They differ from each other but generally adopt the characteristics of the pooled f_m and f_0 shown in figure 4.1. These differences between estimated f_m and f_0 support our assumption about the connection between functional relatedness and transcriptional regulation.

Though the probability as an underlying gene for a few genes can be negative, which has been noticed since the introduction of the procedure (Efron *et al.*, 2001), the maximum of such probabilities for genes in a single eQTL, denoted as $maxP$, seems desirable for most eQTL. Figure 4.1D shows their histogram.

Table 4.1: A singe gene as the most probably gene for large number of traits

gene	max[1]	all[2]	annotation[3]
YNL067W	92	294	Protein component of the large (60S) ribosomal subunit
YNL069C	116	367	N-terminally acetylated protein component of the large (60S) ribosomal subunit
YBR154C	146	436	RNA polymerase subunit ABC27, common to RNA polymerases I, II, and III; contacts DNA and affects transactivation
YNL096C	224	325	Protein component of the small (40S) ribosomal subunit
YOL077C	284	452	Nucleolar protein, constituent of 66S pre-ribosomal particles

[1] number of traits mapped not the gene with the gene as the most probable underlying gene

[2] total number of genes mapped onto the gene

[3] from `http://www.yeastgenome.org/`

When two trait genes have overlapped eQTL, there is chance that the two genes associated with $maxP$ in the two eQTL are a same one. Table 4.1 shows some extreme examples, where a single gene is considered as the most probably underlying gene for a large number of traits. They are unlikely to be explained by coincidence. It can be noticed that most genes listed there are related with ribosome function, which essentially suggests the process of protein synthesis will affect transcript abundance of a large amount of genes. It is quite easy to understand: proteins push yeasts through cell cycles, while cell cycle is tightly associated with genome-wide oscillation of transcriptional activity (Klevecz *et al.*, 2004). Variability in protein production affects cell cycles and hence transcription.

4.4 References

Barabasi, A.-L., and Z. N. Oltvai, 2004 NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION. *Nat Rev Genet* **5**: 101–113.

Brem, R., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752–755.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and

G. de Haan, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

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. *Nat Genet* **37**: 233–242.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher, 2001 Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**: 1151–1160.

Ihmels, J., R. Levy, and N. Barkai, 2004 Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat Biotech* **22**: 86–92.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale, 1999 Multiple Interval Mapping for Quantitative Trait Loci. *Genetics* **152**: 1203–1216.

Klevecz, R. R., J. Bolen, G. Forrest, and D. B. Murray, 2004 A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc Natl Acad Sci U S A* **101**: 1200–5.

Lee, I., S. V. Date, A. T. Adai, and E. M. Marcotte, 2004 A Probabilistic Functional Network of Yeast Genes. *Science* **306**: 1555–1558.

Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble, 2003 Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**: 1275–83.

Luscombe, N. M., M. Madan Babu, H. Yu, M. e. Snyder, S. A. Teichmann, and M. Gerstein, 2004 Genomic analysis of regulatory network dynamics reveals large topological changes. *nature* **431**: 308–312.

Mackay, T. F., 2001 Quantitative trait loci in Drosophila. *Nat Rev Genet* **2**: 11–20.

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.

Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440–5.

Weston, J., A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble, 2004 Protein ranking: from local to global structure in the protein similarity network. *Proc Natl Acad Sci U S A* **101**: 6559–63.

Yvert, G., R. B. 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 rol e of transcription factors. *Nat Genet* **35**: 57–64.

Chapter 5

Prospective Studies

5.1 FDR control in MIM

Let $p_{kj} = \Pr(Q_{kj} = 1|Data)$ for j-th QTL of trait k, where $Q_{kj} = 1$ denotes the QTL is true, $Q_{kj} = 0$ for otherwise. See the introduction chapter for details about this quantity. Given a rejection region and a list of declared QTL, pFDR can be calculated as (Storey *et al.*, 2005)

$$
\begin{aligned}
pFDR &= \frac{\sum_{k=1}^{K} 1 - P_k}{K} && (5.1) \\
P_k &= \prod_{j=1}^{J_k} p_{kj} && (5.2)
\end{aligned}
$$

where K is the number of traits with at least one QTL, J_k is the number of QTLs declared for trait k. Thus, FDR is not for detected QTLs, but for the estimated genetic architecture of traits. We think such FDR estimation procedure agrees with the statistical procedure (sequential genome scan) to find QTL.

Storey *et al.* (2005) showed a way to declare QTL using a sequential genome search while controlling a specific FDR. That is, the rejection region is estimated after performing all sequential tests.

In order to do that, however, they pre-defined a value $J = 2$, the maximal number of QTL for each trait, and did the genome scan for J times for each trait. In each sequential genome scan, they calculated p_{kj} as a function of the maximal likelihood ratio statistic (LRS) across the genome for trait k. Since the LRS for the second QTL is obtained conditional upon the existence of the first QTL, Storey *et al.* (2005) considered p_{k2} as a conditional probability given $Q_{k1} = 1$, or $p_{k2} = \Pr(Q_{k2} = 1|Q_{k1} = 1, Data)$. Thus, after finishing J cycles, P_k, the joint probability that all J

QTL of trait k are true, can be obtained as in formula 5.2. P_k are sorted in descending order. The first K P_k are claimed as significant while controlling FDR at α, where K is the maximal number of P_k such that $pFDR \leq \alpha$ according to formula 5.1. For the yeast data by Brem and Kruglyak (2005), the estimated rejection region included $K = 170$ traits with the minimal $P_k = 0.84$.

Thus, if $p_{kj} < 0.84$ for the j-th QTL of trait k, P_k will also be less than 0.84, and trait k will not reside in the rejection region. This suggests that we can specify a tuning parameter λ, in a way that if $p_{kj} \leq \lambda$ for trait k in the j-th cycle of the genome scan, the joint probability P_k will unlikely be one of the largest K joint probabilities. We will discuss how to specify λ later. Suppose we have already had such a λ. If we have $p_{kj} \geq \lambda$, we include Q_{kj} into the candidate QTL pool. Q_{kj} has not been declared as significant yet. Significance tests will be made after finishing all the genome scan. For $p_{kj} < \lambda$, Q_{kj} is still included in the candidate QTL pool; but no more QTL will be searched for trait k in the following cycles of the genome scan. There are two reasons for this stopping procedure.

- LRS tends to decrease for QTL detected in the later cycle of sequential search. p_{kj}, which is usually a monotonously increasing function of LRS, tends to decrease as well. We generally do not expect that a true QTL's correlation pattern will show up when conditioning on the effect of dubious QTL.
- If the value of P_k is brought down remarkably by a newly added p_{kj}, this would suggest that it is unlikely that trait k will have j QTL. It could be very likely that it has $j-1$ QTL; and it is not sensible to find more QTL.

Introducing λ, appearing to be ad hoc, has the following advantages.

- It restricts the parameter space that we want to explore to find the optimal solution. In the setting of sequential genome scan, introducing λ is an attempt to avoid testing hypotheses that apparently tend to be null, and thus to decrease the proportion of true null hypothesis in all the hypotheses to be tested.
- It provides a stopping rule: stop searching for more QTL of trait k when $p_{kj} < \lambda$. The value of J_k (as in 5.2) is determined in the meanwhile. We do not have to specify J, the maximal number of QTL per trait.

We summarize our proposal to claim as many QTL as possible given a FDR level α in Algorithm 1.

In the algorithm, line 7 ensures that if the j-th QTL for trait k is claimed as significant, all previous $j-1$ QTLs are also included. LRS and p_{kj} generally decrease as j increases. So line 7 is designed to deal with exceptions. Line 18 is designed to deal with the cases that the specified λ is too stringent. This step allows a grid search for the optimal λ, starting from a certain large value. $\Delta\lambda$ can be 0.05 or smaller, depending on the computational burden to do additional genome scans and to estimate the non-linear function between LRS and p_{kj}.

In the QTL analysis for this yeast population, we can set $\lambda = 0.8$ from inspecting Storey *et al.* (2005)'s results. For a general case, if we want to control FDR $\leq \alpha$, we can set the initial value of λ at $1 - \alpha$.

5.2 Borrow testing power from correlated traits

Lan *et al.* (2006) noticed that many transcripts with similar biological annotations

Algorithm 2 FDR control in MIM

```
1: Obtain the null distribution of genome-wise maximal LRS using a permutation test.
2: Forward selection of QTL main effects as in MIM. For $Q_{kj}$, compare its LRS to the null distribution to get $p_{kj}$, and find all $p_{kj} \geq \lambda$ .
3: Sort all $p_{kj}$ from largest to smallest such that $p_{kj}^{(1)} \geq p_{kj}^{(2)} ... \geq p_{kj}^{(m)}$ (suppose that $m$ LRS are collected)
4: Let $\Gamma$ be a rejection region, which is to be determined in the following steps.
5: for $i = 1$ to $m$ do
6:   include $p_{kj}^{(i)}$ in $\Gamma$
7:   for $j' = 1$ to $j - 1$ do
8:     if $p_{kj'}$ is not in $\Gamma$ then
9:       include $p_{kj'}$ in $\Gamma$
10:    end if
11:  end for
12:  calculate pFDR as formula 5.1 given $\Gamma$
13:  if pFDR $> \alpha$ then
14:    remove all $p_{kj}$ included in this cycle (starting from line 5) from $\Gamma$
15:    exit iteration
16:  end if
17: end for
18: if all $m$ $p_{kj}$ are in $\Gamma$ and pFDR $< \alpha$ then
19:   decrease $\lambda$ by $\Delta\lambda$ and go to line 2 of the algorithm to collect more $p_{kj}$
20: end if
```

have linkage statistic peaks at the similar genomic location, though many peaks are below the significant threshold. Such an observation is a perfect example for the joint analysis of multiple genetic related traits.

The difference between the joint analysis of multiple traits and the single trait analysis is that, a more complex model is used in the former approach: correlation structure of multiple traits are explicitly modelled, and thus, data of these multiple traits can be analyzed together. Such an approach could potentially increase the statistical power of QTL detection and also the accuracy of QTL localization.

It is impractical to expand this approach to thousands of traits as in the case for eQTL mapping. An alternative approach is to fit a multiple trait model for a smaller cluster of correlated traits (Lan *et al.*, 2003), however, this requires a procedure to group traits. It is also possible to do dimension reduction using principle components (Lan *et al.*, 2003), however, it will be hard to interpret QTL for such composite traits.

Here, I suggest an eQTL detection method in which

- each trait is analyzed separately
- information is borrowed from all traits with correlated expression profiles in the sample
- correlation structure among trait phenotypes is not explicitly estimated

Kendziorski *et al.* (2006) suggested a new angle of expression QTL mapping by borrowing statistical research effort in detecting differential expressed (DE) genes from microarray data: instead of identifying QTLs for each expression trait one by one, a 'marker based' approach is to test markers one by one for DE among sample subgroups fixed with different genotypes at the marker. It is not a simple change in the order of testing. The key point is that many DE test procedures use information of the complete set of expression profiles when testing the DE of a single trait. Optimal discovery procedure (ODP) (Leek *et al.*, 2006) is such a test.

Assuming each marker has G types of genotypes, we can number each genotype from 1 to G. Let $M_{kj} = g, g = 1, 2, ..G$ denote that the k-th marker of individual j has the genotype g. Let y_{ij} be the gene i's transcript abundance of individual j, $i = 1..m, j = 1..n$. The expression profile of a single gene is written as $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})$. Define a set J_{kg} as all the j such that $M_{kj} = g$, i.e, the subset of individuals who have genotype g on the k-th marker. Thus, $\mathbf{y}_i = (\mathbf{y}_{ig}), g = 1..G$, where $\mathbf{y}_{ig} = (y_{ij})_{j \in J_{kg}}$,i.e., a partition of $\mathbf{y}_i$ according to individuals' genotype at M_{kj}.

We will assume an independent normal distribution for each y_{ij} given a trait i. Under the null hypothesis that all y_{ij} have a same mean, or no differential expression among $\mathbf{y}_{ig}$, we can write out the likelihood function as

$$L(\mathbf{y}_i; u_{i0}, \sigma_{i0}^2) \propto \exp[-\frac{\sum_{j=1}^{n}(y_{ij} - u_{i0})^2}{2\sigma_{i0}^2}] \tag{5.3}$$

Under the alternative hypothesis that given the k-th marker, individuals in different J_{kg} have different mean expression level of gene i: u_{ig}, but the same variance σ_{i1}^2, the likelihood function for the data is

$$\prod_g L(\mathbf{y}_{ig}; u_{ig}, \sigma_{i1}^2) \propto \exp[-\frac{\sum_{g=1}^{G}\sum_{j \in J_{kg}}(y_{ij} - u_{ig})^2}{2\sigma_{i1}^2}] \tag{5.4}$$

Following Storey (2005), we can calculate t_{hk}, the ODP statistics for expression trait h on marker k as:

$$t_{hk} = \frac{\sum_{i=1}^{m} \prod_{g} L(\mathbf{y}_{hg}; \hat{u}_{ig}, \hat{\sigma}_{i1}^2)}{\sum_{i=1}^{m} L(\mathbf{y}_h; \hat{u}_{i0}, \hat{\sigma}_{i0}^2)} \tag{5.5}$$

where

$$\begin{aligned}
\hat{u}_{i0} &= \sum_{j=1}^{n} y_{ij}/n && (5.6)\\
\hat{u}_{ig} &= \sum_{j \in J_{kg}} y_{ij}/n_g, n_g \text{ is the length of } J_{kg} && (5.7)\\
\hat{\sigma}_{i0}^2 &= \sum_{j=1}^{n} \frac{(y_{ij} - \hat{u}_{i0})^2}{n-1} && (5.8)\\
\hat{\sigma}_{i1}^2 &= \sum_{g=1}^{G} \sum_{j \in J_{kg}} \frac{(y_{ij} - \hat{u}_{ig})^2}{n-G} && (5.9)
\end{aligned}$$

Storey (2005) suggested that y_{ij} be centered around zero within each trait to ensure 5.5 is statistically justifiable. Thus, $\hat{u}_{i0} = 0$.

It should be noted that in formula 5.5, we insert $\mathbf{y}_h$ into the likelihood functions of each expression trait under both null and alternative hypotheses. That is the unusual feature of ODP. However, in this way, any gene i with similar variation patterns as gene h will contribute substantially to the ODP statistic of gene h. From formula 5.3 and 5.4, it is also clear that the contribution from gene i is weighted by its variance under null and alternative hypothesis.

In 5.5, likelihood functions for different expression traits are summed together. ODP statistic avoids explicitly modelling the correlation structure for multiple traits, which can be over thousand traits in eQTL mapping settings, but is still able to borrow statistical power from correlated traits.

ODP statistic is explicitly designed to maximize the expected number of true positive declarations with each expected false positive inference, and thus turns out to be much powerful than other methods (Leek *et al.*, 2006; Storey, 2005).

Based on ODP statistic t_{hk}, we can perform a forward selection for multiple markers (QTL) co-segregating with trait h.

1. Calculate t_{hk} for all traits across the genome.

2. Suppose that for trait h, t_{hk} is maximized at marker $M(h1)$ in the 1st cycle of the sequential search. Record the maximum as $t_{hM(h1)}$. Perform the analysis for all traits.

3. For each trait h, regress phenotypic values onto $M(h1)$, and replace the value of y_{hj} with its corresponding residual in the regression. Center these new y_{hj} around zero so that $\sum_j y_{hj} = 0$. Go back to step 1 to find $M(hi), i = 2, 3..$, the marker where the ODP statistic for trait h is maximized in cycle i of the sequential genome. Record the corresponding $t_{hM(hi)}, i = 2, 3..$

4. Stopping rules for the sequential genome scan based on likelihood ratio statistics, or F statistics can be applied to ODP statistics.

    - We can follow Storey *et al.* (2005) by restricting ourselves in all 2-QTL models. $t_{hM(hi)}, i = 1, 2$ are then compared with their null distributions (from permutation tests) to declare co-segregating markers while controlling FDR at certain level.
    - We can fix a type I error rate for genome-wise scan of a single trait and obtain the corresponding threshold for $t_{hM(hi)}$ through permutation tests (Churchill and Doerge, 1994; Doerge and Churchill,

1996). The sequential search for QTLs of trait h stops when the first time $t_{hM(hi)}, i = 1, 2, 3...$ is below its threshold. pFDR, the average probability of declaring a false multiple QTL model for a trait can be obtained as in 5.1.

5.3 References

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–71.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**: 285–94.

Kendziorski, C., M. Chen, M. Yuan, H. Lan, and A. AD., 2006 Statistical Methods for Expression Quantitative Trait Loci (eQTL) Mapping. *Biometrics* **62**: 19–27.

Lan, H., M. Chen, J. B. Flowers, B. S. Yandell, D. S. Stapleton, C. M. Mata, E. T. Mui, M. T. Flowers, K. L. Schueler, K. F. Manly, R. W. Williams, C. Kendziorski, and A. D. Attie, 2006 Combined Expression Trait Correlations and Expression Quantitative Trait Locus Mapping. *PLoS Genet* **2**: e6.

Lan, H., J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell, and A. D. Attie, 2003 Dimension Reduction for Mapping mRNA Abundance as Quantitative Traits. *Genetics* **164**: 1607–1614.

Leek, J. T., E. Monsen, A. R. Dabney, and J. D. Storey, 2006 EDGE: extraction and analysis of differential gene expression. *Bioinformatics* **22**: 507–508.

Storey, J., 2005 The optimal discovery procedure: a new approach to simultaneous significance testing. UW Biostatistics Working Paper Series Working Paper, 259.

Storey, J. D., J. M. Akey, and L. Kruglyak, 2005 Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.