Statistics and Its Interface Volume 4 (2011) 389–402

A combined p-value approach to infer pathway regulations in eQTL mapping

Shaoyu Li Barry L. Williams and Yuehua Cui*

The genetic bases of complex traits often involve multiple inherited genetic factors that function in a network basis. By promoting or reducing the expression of functional genes that are directly or indirectly related to a trait, gene regulation has been proposed as a major determinant of trait variation. The combined analysis of genetic and gene expression data, termed genetical genomics analysis or eQTL mapping, holds great promise in disentangling the mechanism of gene regulation. Given that genes function in a network basis, the detection of a genetic system as a whole could shed novel light into the role of gene regulation. We hypothesized that gene expression changes are often caused by the regulation of a set of variants that belongs to a common genetic system (e.g., a gene network or a pathway). We proposed to combine individual signals (e.g., p-values) within a genetic system to form an overall signal while considering correlations between variants, with the goal of inferring the role of the whole system in regulating gene expression in an eQTL mapping framework. A Satterwhite's approximation method is applied to approximate the distribution of the combined p-values. Both simulation and real data analysis showed the relative merits of the combined method. Our method provides a novel strategy in addressing questions related to gene regulations from a systems biology perspective.

Keywords and phrases: Gene regulation, Gene network, Genetical genomics, Pathway regulation, Satterwhite's approximation, Systems biology.

1. INTRODUCTION

Advancements in microarray, genotyping and next generation sequencing technologies have made it possible to measure thousands of gene expression profiles simultaneously, and to genotype thousands of genetic markers, in order to understand the function of a living organism in a systematic way. The integrative analysis of these two sources of biological information, termed expression quantitative trait loci (eQTL) mapping or genetical genomics analysis, holds great promise in elucidating the genetic architecture of gene expression and gene regulation (Jansen and Nap 2001; Schadt et al. 2003). In a typical eQTL mapping study, each gene expression level is considered a single trait, and the goal of such studies is to identify the genetic loci that mediate expression changes on a genome-wide scale. The testing unit is often a single gene against a single marker (e.g., a SNP), the so called single marker – single trait analysis. Given that the expression of a gene or a network of genes may be regulated by a group of variants functioning together as a system, studying gene regulations by focusing on the joint function of variants in a system could shed novel light into the complexity of a biological system. There are quite a few review articles for eQTL mapping in the literature. For example, readers are referred to Kendziorski and Wang (2006) for a review of statistical methods in eQTL mapping, and to Gilad et al. (2008) and Li and Burmeister (2005) for a general review of eQTL mapping studies. It is not the focus of this paper to given an exhaustive review of the commonly used methods in eQTL mapping, to list a few recent methodological developments, see for example Chun and Keles (2009).

It is commonly recognized that genes in a pathway or network act in a coordinated manner to fulfill a joint task. Thus analysis from a systems biology perspective, for instance, focusing on genetic variants in terms of pre-defined pathways/networks, can provide valuable biological insights into gene function and regulation, which otherwise can not be easily achieved by single marker–single trait analysis (Mootha et al.; 2003,Wessel et al.; 2007, Lee et al. 2007; Wu et al. 2008). Moreover, variants in a genetic pathway often confer moderate effects in mediating the expression change of a gene or a gene network, which makes it difficult to detect individual effect and consequently leads to low power in single marker analysis. From a biological point of view, signals in a genetic system, even though individually not significant (say p-values of 0.06 for an extreme case), many such values for related genes within a pathway or network when taken together may suggest the relative importance of that particular genetic system in mediating gene expression changes. By a genetic system we mean a group of genes within a genetic functional category which can be obtained from various sources such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004), Gene Map Annotator and Pathway Profiler (Dahlquist et al. 2002) and Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) (Thorn et al. 2005), or a category of multiple loci defined by SNP physical locations.

The above thoughts motivated us to consider a joint analysis in which multiple signals are combined together to indi-

*To whom the correspondence should be addressed

cate the contribution of the overall system. Herein, we argue that a joint analysis could provide additional insights into gene function and regulation that otherwise could not be achieved by looking at individual signals alone. We propose to combine individual p-values in a genetic system (e.g., KEGG category) while considering the correlations among them, to form an overall signal for inference of shared gene expression patterns in an eQTL mapping framework.

Methods of combining p-values have been applied to a wide range of problems, including genome-wide association studies (e.g., Peng et al., 2010; Yu et al., 2009), multiple endpoints studies in clinical trails and meta-analysis, and detecting differentially expressed genes (Hess and Iyer, 2007). There are different p-value combination methods in the literature, for example, the Fisher's combined p-value approach (Fisher, 1932); the truncated product method (Zaykin et al., 2002); the rank truncated product method (Dudbridge and Koeleman, 2003); and the weighted truncation product method (De la Cruz et al. 2010). A commonality among these combining methods is to first take a transformation of individual p-values and then evaluate the distribution of the combined statistic. However, when individual tests are not independent, the distribution of the combined statistic is difficult to obtain. Moreover, there is no analytical criterion for choosing the truncation threshold for the truncated product methods.

For multiple individual tests in a genetic system, it is known that they are not independent due to linkage disequilibrium (LD) or functional interactions between variants. Regarding the concern of correlations among individual tests, methods that ignore correlations and treat them independently will obviously affect the accuracy of the results and could lead to either inflated false positives or false negatives. Some work has been done to handle correlations when combining individual p-values. For example, one could estimate the empirical null distribution of the combined statistic by a simulation-based procedure (Zaykin et al., 2002); approximate the null distribution based on a known correlation matrix (Kost and McDermott, 2002); or apply the most widely used permutation approach. Although, permutation approaches, when performed appropriately, provide an unbiased estimation of the null distribution and are widely considered the gold standard with which other tests are compared, their main disadvantage is the computational cost (Conneely and Boehnke, 2007). For example, to get an empirical p-value of 10^{-5}, at least 10^5 permutations are needed.

When a large number of tests are involved in a study, alternative methods that can provide similar accuracy would be attractive. Brown (1975) proposed to combined dependent tests assuming a multivariate normal distribution of the test statistics with a specified covariance structure. The method later on was extended by Kost and McDermott (2002) assuming a known covariance matrix up to a scaler quantity. The assumption of a known covariance matrix limits their application as in most cases the distribution of the test statistic is unknown with an unknown covariance matrix. In this article, we focused our attention on the Fisher's combination statistic and proposed to approximate its null distribution with a scaled chi-square distribution while considering correlations among individual tests. We proposed different strategies to estimate the correlation information. The rest of the paper is organized as follows. Section 2 discusses the approximation evaluated with the Satterthwaite's approximation. Section 3 provides simulation studies to evaluate the accuracy, type I error rate and power of the approximation method. Section 4 applies the method to a yeast eQTL mapping data set to identify pathway regulation patterns, followed by the discussion in section 5.

2. STATISTICAL METHODS

2.1 Pattern of gene regulation

It has been commonly recognized that gene regulation plays a pivotal role in determining trait variation in natural populations by promoting or reducing the expression of functional genes that are (in)directly related to a phenotypic trait. Thus, using gene expression values as phenotypes in eQTL mapping can provide additional insights into gene regulation, particularly in distinguishing *cis*- and *trans*-regulation that is associated with trait variation, which otherwise can not be decerned in a traditional QTL mapping study (Alberts et al. 2005). *Cis*-acting eQTLs are defined as sequence variants that are located within or close to the gene being regulated and hence are attractive candidate genes for functional QTLs mapped to the same location (Hubner et al., 2005). *Trans*-regulated eQTLs are those remotely located away from the gene being regulated, tending to cluster together and sharing similar regulatory mechanisms that can be used to identify gene clusters within the same pathways (Mueller et al. 2006; Petretto et al. 2006). Given that genes function in networks, the identification of regulatory elements, as well as the "master regulators" that affect the expression of hundreds of genes, can greatly enrich our understanding of gene regulatory networks, and ultimately help us gain novel insights into the genetic architecture of complex traits (Yvert et al. 2003; Petretto et al. 2006).

Figure 1 shows several possible gene regulation patterns. Figure 1(A) and 1(B) show *cis*- and *trans*-regulation patterns, respectively. Figure 1(C) indicates that the same gene can be regulated by multiple *trans*-regulatory loci. Each of these regulatory loci are associated with specific genetic variants. In the context of eQTL mapping, we are trying to identify genetic variants that are associated with these regulatory changes and likely regulate gene expression. To map eQTLs as illustrated in Figs. 1(A)-(C), single marker–single trait analysis can be applied followed by multiple testing corrections. Figure 1(D) shows that a regulatory element can regulate multiple genes, among which some share a common network. When multiple gene expressions are grouped into

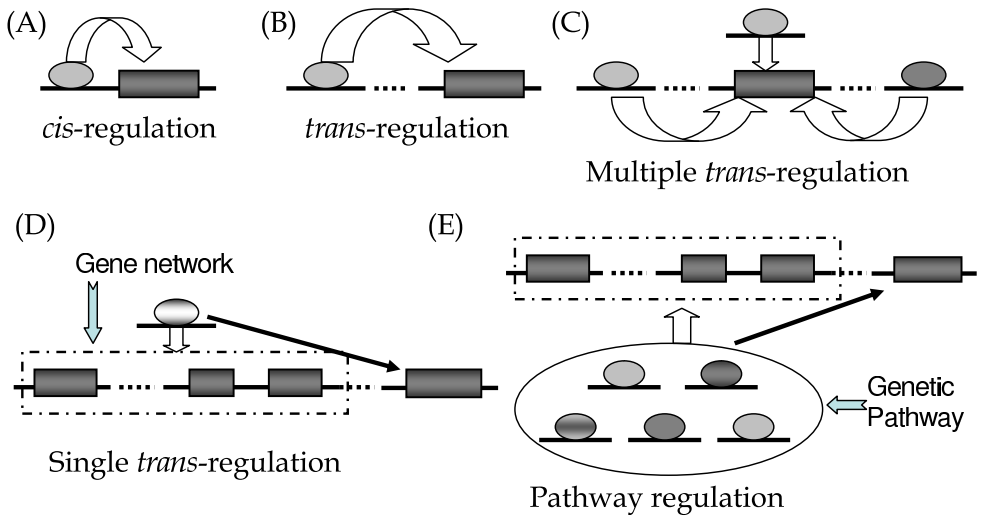

Figure 1. Various patterns of gene regulation: (A) *cis*-element regulates its own gene expression; (B) *trans*-element regulates downstream gene expression; (C) multiple *trans*-elements regulate the same gene expression; (D) single *trans*-element regulates single gene expression or multiple gene expressions in a network (i.e., gene network); and (E) multiple regulators in a genetic pathway function jointly to regulate multiple gene expressions in or not in a network. The shaded ovals and rectangular represent regulatory elements and coding genes, respectively. The dotted lines imply that genes are located in different regions.

a network or a pathway, the identified regulators are termed as network or pathway regulators, and methods for this purpose have been developed (e.g., Li et al. 2010). Figure 1(E) shows that the expression of a single gene or a network of genes is regulated by the joint function of multiple genetic variants, potentially belonging to a common genetic system (e.g., a genetic pathway). The signal perturbation of a genetic system could cause the expression change of a gene or a network of genes, and consequently result in phenotypic changes such as a disease. In this work we focused our analysis in identifying pathway regulation as shown in Figure 1(E). The identification of pathway regulations would help us better understand the genetic architecture of gene expression and regulation from a systems biology perspective.

2.2 The Satterwhite's approximation

As we mentioned in the introduction section, a genetic system can be defined as a genetic pathway from the KEGG database or a GO term, or as a group of variants located physically close to each other. We hypothesize that the signal perturbation of a genetic system could lead to the expression change of a single gene or a network of genes. We assume there are L SNP variants in a given genetic system. For the L SNPs, we conduct L individual tests and obtain L individual test statistics or p-values. Depending on the number of genotype categories at each locus and the expression phenotype distribution, different tests can be applied. For example, a two-sample t-test or Hotelling's T^2 test can be applied depending on whether the response is a single gene expression value or multiple gene expression values, while assuming there are two possible genotype categories at a locus (e.g., in a recombinant inbred line or yeast population). We tried to combine individual signals in a genetic system to determine if it, as a whole system, underlies the expression changes of genes, and hope to gain novel insights into gene regulations from a systems biology perspective.

Let $p_1, p_2, \cdots, p_L$ be the p-values for L individual two-sided tests, $H_{i,0} : \mu_{i1} = \mu_{i0}$ versus $H_{i,1} : \mu_{i1} \neq \mu_{i0} (i = 1, 2, \cdots, L)$ assuming there are two genotype categories (denoted as 1 and 0) at each locus. Define $z_i = -2 \log p_i$. Under the null hypothesis of no genetic effect, each of the L p-values is uniformly distributed and $z_i \sim \chi^2_2$ for $i = 1, \cdots, L$. If we assume the L tests are independent, the Fisher's combined statistic $T = \sum_{i=1}^{L} z_i \sim \chi^2_{2L}$ under the global null hypothesis of no genetic effect.

When multiple genetic variants are considered as a system, they are more or less correlated. Thus the L p-values are not independent and the Fisher's chi-square distribution with $2L$ degrees of freedom (d.f.) does not hold. Here we proposed to approximate T by a scaled chi-square distribution under the null by applying the Satterwhite's approximation method. We assume that the combined statistic T follows a scaled chi-square distribution, i.e.,

$$T = \sum_{i=1}^{L} z_i \dot{\sim} a\chi^2_g. \tag{1}$$

The scale parameter a and the d.f. g are chosen so that the first and second moments of the scaled chi-square distribu-

tion and the distribution of T under the null are identical. For correlated p-values, the expectation and variance of the statistic T under the null can be obtained as

$$E(T) = E(\sum_{i=1}^{L} z_i) = 2L,$$

$$\begin{aligned} Var(T) &= Var(\sum_{i=1}^{L} z_i) \\ &= \sum_{i=1}^{L} Var(z_i) + 2\sum_{j<i} Cov(z_i, z_j) \\ &= 4L + 8\sum_{j<i} \rho_{ij}, \end{aligned}$$

where ρ_{ij} is the correlation between the log-transformed p-values z_i and z_j.

By equating the first and the second moments of T and $a\chi_g^2$, we have

$$E(a\chi_g^2) = ag = E(T) = 2L,$$

and

$$Var(a\chi_g^2) = 2a^2 g = Var(T) = 4L + 8\sum_{j<i} \rho_{ij}.$$

Solving the two equations, we obtain

$$\hat{a} = \frac{4L + 8\sum_{j<i} \rho_{ij}}{4L} = 1 + \frac{2\sum_{j<i} \rho_{ij}}{L}, \tag{2}$$

$$\hat{g} = \frac{2L}{\hat{a}} = \frac{2L^2}{L + 2\sum_{j<i} \rho_{ij}}. \tag{3}$$

When the L SNPs are completely independent, i.e., $\rho_{ij} = 0 \; \forall \; i, j$, it can be seen that the approximation is the same as the distribution of the Fisher's combined statistic assuming independence. When the L SNPs are completely dependent, i.e., $\rho_{ij} = 1 \; \forall \; i, j$, then $\hat{a} = L$ and $\hat{g} = 2$. In this case, the statistic T is just a sum of L independent χ_2^2 variables. For $-1 < \rho_{ij} < 1$, parameters a and g approximate the distribution of T, where a and g can be estimated by Equations (2) and (3). In reality, we rarely see negative correlations for a two-sided test. So the restriction of $2\sum_{j<i} \rho_{ij} > -L$ to get positive estimates of a and g is easily met. The challenge remaining is to estimate the correlation between z_i and z_j from the data. In the following, we illustrate how to estimate the correlation ρ_{ij}.

2.3 Estimating the correlation matrix

Let $\boldsymbol{z} = (z_1, \cdots, z_L)$ be a vector of log-transformed p-values and let Γ be the correlation matrix of $\boldsymbol{z}$. From the above approximation we can see that the accuracy of the approximation to the distribution of T depends largely on how well the correlation matrix Γ is estimated. Assuming a multivariate normal distribution of the test statistics, Brown (1975) proposed to estimate Γ with a completely specified covariance matrix. The author argued that the covariance between z_i and z_j is a function of the correlation between the ith and jth variables under the group of affine transformation. This is however not true in a genetic study, and there is no analytically closed form for the structure of Γ. In this paper, we propose two methods to approximate Γ, which are detailed in the follows.

2.3.1 Estimating the correlation matrix by permutation

Since we want to approximate the null distribution of T, we need the correlation matrix of the transformed p-value vector $\boldsymbol{z}$ under the null hypothesis. Permutation was applied to generate random samples of $\boldsymbol{z}$ by reshuffling the relationship between the gene expression values and genetic markers, where genetic variants for each individual in a system are maintained as a vector to preserve their correlation structure. For each permutation, we would have a vector of p-values, $p^b = (p_1^b, p_2^b, \cdots, p_L^b)$ and also the transformed p-values $\boldsymbol{z}^b = (z_1^b, z_2^b, \cdots, z_L^b)$. The correlation matrix for $\boldsymbol{z}$ under the null then can be estimated by the sample covariance of the permuted random sample: $\boldsymbol{z}^b (b = 1, 2, \cdots, B)$, and B is the total number of permutations (say 1000). The sample correlation matrix obtained from the permuted samples were used as the estimate of Γ. No assumption is required for the distribution of the test statistics at this step. Generally speaking, the larger the data dimension (L), the more the permutations are required.

2.3.2 Estimating the correlation matrix by LD approximation

Note that multiple variants in a genetic system are either physically close to each other or functionally correlated. The correlation information is more or less reflected by LDs between the variants. This motivates us to approximate Γ by LDs among SNP variants whose individual p-values are to be combined. Unfortunately there is no analytical solution to assess the relationship between the correlations of z and the LDs. We checked the relationship between the LDs of SNP variants (measured by R^2) and the correlation structure of $\boldsymbol{z}$. To begin with a simple example, we considered two SNP variants, each with a minor allele frequency (MAF) of $q = 0.1$ $(0.3, 0.5)$. For a given MAF, the range of LD denoted by D is given by

$$max\{-q_1 q_2, -(1-q_1)(1-q_2)\} \leq D \leq min\{q_1(1-q_2), q_2(1-q_1)\},$$

where q_1 and q_2 denote the MAF for SNPs at two different loci. If we assume the same MAF for both SNPs, the range of D becomes $max\{-q^2, -(1-q)^2\} \leq D \leq q(1-q)$ and the range of $R = \frac{D}{\sqrt{(q_1(1-q_1)q_2(1-q_2))}} = \frac{D}{q(1-q)}$ is $max\{-q/(1-q), -(1-q)/q\} \leq R \leq 1$.

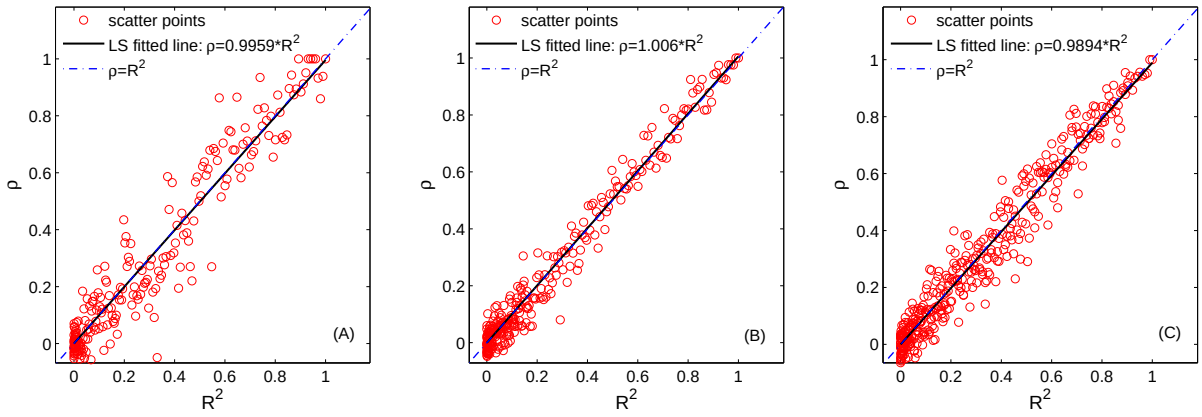

Figure 2. Scatter plots of correlation coefficient ρ vs LD R^2. The blue line is $\rho = R^2$, black line is the least square fitted line. (A) $MAF = 0.1$, fitted function: $\rho = 0.996R^2$; (B) $MAF = 0.3$, fitted function: $\rho = 1.006R^2$; (C) $MAF = 0.5$, fitted function: $\rho = 0.99R^2$.

For a fixed MAF, we generated genotypes for two SNPs with different values of D (hence R) in a given range (following the procedure described in the LD-based simulation section). Phenotypes were simulated independent of the SNPs (i.e. under the null distribution) and then tested for association between the phenotype and the two SNP markers with p-values denoted by p_1 and p_2. For a given R value, the correlation coefficient of the two transformed p-values $z_1 = -2\log p_1$ and $z_2 = -2\log p_2$ was calculated from 1000 simulated samples. Scatter plots of the correlation coefficient ρ against R^2 corresponding to MAF 0.1, 0.3 and 0.5 are given in Figure 2. The three plots clearly indicate a linear relationship between ρ and R^2. The least squares fitted lines (black) almost perfectly overlap with the $\rho = R^2$ lines (blue). We also tried various allele frequency combinations for the two SNPs and found very similar relationships. Since a two-sided test was performed, even with negatively correlated SNPs, their p-values are still positively correlated. This explains why we rarely see negative correlations between the log-transformed p-values. We assessed the relationship for a real eQTL data set applied in this study (discussed in the real data analysis section). A similar relationship was also observed (Figure 3). The assessment in simulation and real data indicates that R^2 provides a good approximation to the correlation between the log-transformed p-values.

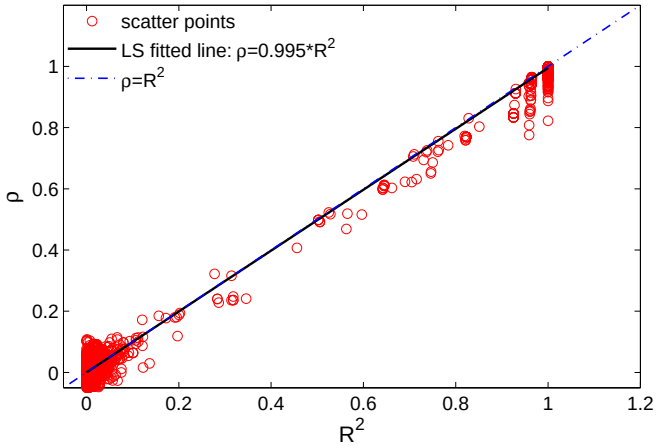

Figure 3. Scatter plot of the correlation coefficient ρ and R^2 for the YEAST eQTL data set. The black line is the least square fitted line: $\rho = 0.995R^2$ and the blue line is a straight diagonal line.

3. SIMULATION STUDY

3.1 Accuracy of the scaled χ^2 approximation

The accuracy of the scaled chi-square approximation was evaluated by a χ^2-plot. Considering two p-values, p_1 and p_2, which are correlated with $\text{corr}(p_1, p_2) = \rho*$. We generated 1000 random samples of p-values with a given correlation $\rho*$. The corresponding combined statistic T for the 1000 simulated samples were obtained. The estimated correlations between log-transformed p-values were then used to estimate a and g. Figure 4 plots the approximated percentiles using $\hat{a}\chi^2_{\hat{g}}$ (right panel) and χ^2_{2L} (left panel) versus the observed empirical percentile of T. As shown in the figure, points of percentiles of scaled chi-square distribution and the empirical percentiles lie roughly on a straight line, while χ^2-plot for the χ^2_{2L} approximation deviates from the straight line, especially at the tail. The plots demonstrate that the scaled chi-square distribution provides a much more accurate approximation to the distribution of T under the null than a regular chi-square distribution does. Simply ignoring the correlations among the test statistics would result in biased approximation and wrong inference.

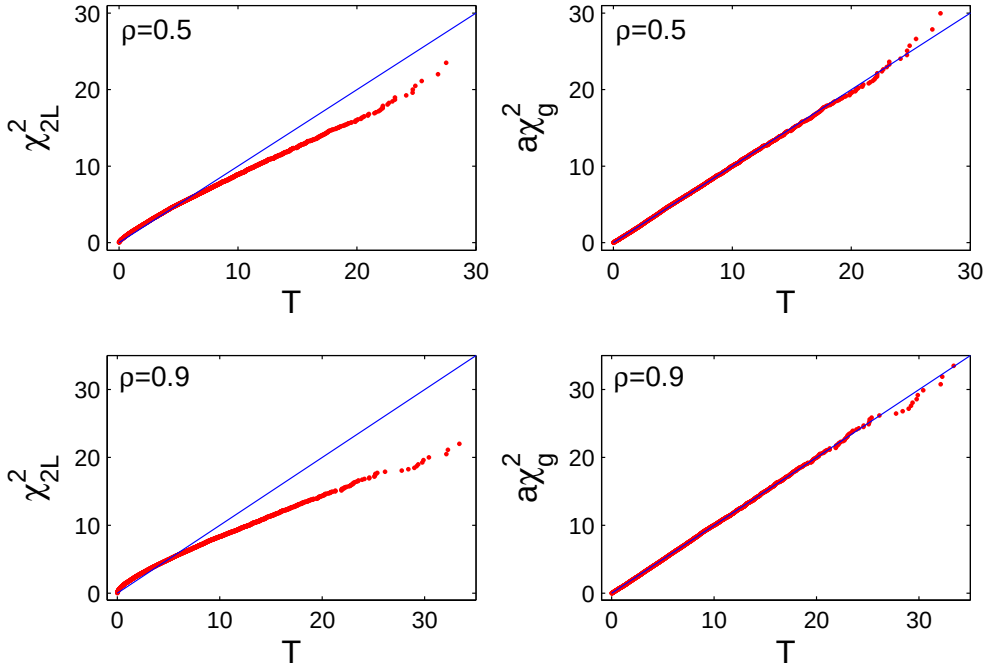

Figure 4. χ^2 plot for percentiles of the observed statistic T against the χ^2_{2L} approximation (left panel) and $a\chi^2_g$ approximation (right panel). Two correlations were assumed: $\rho = 0.5$ (upper panel) and $\rho = 0.9$ (lower panel).

3.2 Simulation design

3.2.1 Genotype simulation

We simulated genotypes for one genetic pathway with multiple SNP variants. These variants function together as a whole system to regulate expression changes of a single gene or a network of genes. Two methods were used to simulate the genotype data. The first method, termed LD-based simulation, generates SNP genotype data based on pairwise LD structure. The second method is a real data-based simulation which mimics gene structure and LD patterns of a real data set by sampling genotypes directly from the data.

LD-based simulation: Let q_A and q_B be the frequencies of two alleles A and B for two adjacent SNPs, with LD denoted by D. The frequencies of four haplotypes can be expressed as $p_{ab} = (1-q_A)(1-q_B)+D, p_{AB} = q_Aq_B+D, p_{Ab} = q_A(1-q_B) - D, p_{aB} = (1-q_A)q_B - D$. Assuming HardyWeinberg equilibrium, the SNP genotype at locus A can be simulated assuming a binomial distribution. Locus B can be simulated conditional on locus A with the conditional probability given by

$$P(B|A) = \frac{P(BA)}{P(A)} = \frac{p_{AB}}{q_A} = \frac{q_Aq_B + D}{q_A}. \tag{4}$$

This illustration is for simulating a haploid genome (e.g., yeast). The same idea can be applied to simulate a diploid genome. The advantage of this simulation strategy is that we can easily control the pairwise LD pattern between adjacent SNPs. We assume genes in a pathway are in linkage equilibrium (The assumption is not required for the method, but is used only for illustration of the feasibility of the proposed approach to different applications). SNPs within each gene are in LD and the genotypes for SNPs in each gene were simulated by the LD-based simulation approach. We simulated SNP genotypes for four individual genes, $G1(8)$, $G2(5)$, $G3(3)$ and $G4(4)$, where the number in parenthesis indicates the number of SNP markers in the corresponding genes. The four genes were assumed to belong to one genetic pathway. LDs for SNPs within each gene were set to $R^2 = 0.9$.

Real data-based (RD) simulation: To simulate SNPs which mimic the gene structure and LD patterns among SNPs in a real genetic pathway, we took genotype vectors for SNPs within the #20 genetic pathway ("00290", Valine, leucine and isoleucine biosynthesis) in the yeast data set. Genotype vectors were randomly drawn with replacement from the real data to create a simulation sample. This genetic pathway has four individual genes with 14 SNPs in total. Missing genotypic values were imputed before the random draw. We found that the pairwise LDs in this pathway varies with $D \in (-0.035, 1)$ and $R \in (-0.14, 1)$.

3.2.2 Phenotype simulation

Several simulation scenarios assuming different gene actions were considered (Table 1). Model I considers the case

Table 1. List of data generating models

Model	Gene action
I	$y = \mu + \epsilon$
II	$y = \mu + \beta_1 S_1 + \beta_2 S_2 + \beta_7 S_7 + \beta_8 S_8 + \epsilon$
III	$y = \mu + \beta_1 S_1 + \beta_2 S_2 + \beta_{15} S_1 S_5 + \beta_{38} S_3 S_8 + \epsilon$
IV	$y = \mu + \beta_1 G_{1,1} + \beta_2 G_{1,2} + \beta_3 G_{2,1} + \beta_4 G_{2,2} + \beta_5 G_{3,2} + \beta_6 G_{1,3} G_{3,2} + \epsilon$
V	$y = \mu + \beta_1 G_{1,1} + \beta_2 G_{1,2} + \beta_3 G_{2,1} + \beta_4 G_{2,2} + \beta_5 G_{2,2} G_{2,3} + \beta_6 G_{1,5} G_{4,4} + \epsilon$

Where S_j represents the jth SNP in a genetic pathway; $G_{i,j}$ represents the jth SNP in the ith gene. The effect of β_{ij}'s were considered the same.

in which there is no genetic effect at all. So model I is the null model we used to assess the false positive rate. Model II assumes only main SNP effects in a genetic pathway (SNPs 1, 2, 7 and 8). Model III assumes main SNP effects (SNPs 1 and 2) as well as the interactions between SNPs 1 and 5 and between 3 and 8. Model IV and V simulate phenotypes considering the gene structure in a genetic pathway. Interactions were considered for SNPs in different genes. Model IV considers interactions only when the corresponding gene has a main effect. Model V assumes there is an interaction effect between two genes and one of which has no marginal main effect.

We applied model II and model III to simulate phenotypes with genotype simulated by the RD-based simulation method. The LD-based simulation method were applied for model IV and model V to generate phenotype data. Thus four different simulation scenarios were considered: (A) RD-based genotype + Model II phenotype; (B) RD-based genotype + Model III phenotype; (C) LD-based genotypes + Model IV phenotype; and (D) LD-based genotypes + Model V phenotype. Type I error rate was assessed with phenotypic data simulated by Model I.

3.3 Simulation Results

We evaluated the type I error rate and power of the scaled chi-square approximation to infer genetic regulatory patterns. The type I error rate was estimated by simulating 1000 data sets under the null distribution (Model I). Similarly, we estimated power by simulating 1000 data replicates for each model (Model II-V). Two-sided two sample t-tests were applied to test for associations between SNP markers and a quantitative trait y. Individual p-values for all SNP markers within the pathway were then combined to form the test statistic $T = -2\sum_{i=1}^{L} \log p_i$. For each simulated data set, a p-value for the combined statistic T is assessed and is denoted by $p^c_{\chi^2_{2L}}$, $p^c_{a\chi^2_g}(\text{perm})$, $p^c_{a\chi^2_g}(R^2)$ and p^c_{perm}. For $p^c_{\chi^2_{2L}}$, the combined p-value follows a χ^2_{2L} distribution under the null; for $p^c_{a\chi^2_g}(\text{perm})$ and $p^c_{a\chi^2_g}(R^2)$, the combined p-value follows a scaled $a\chi^2_g$ distribution, where parameters a and g were estimated by using correlations approximated by the permutation-based and the LD-based approximation (i.e., $\rho = R^2$) approaches, respectively; and for p^c_{perm}, the significance of the combined p-values were assessed by permutation tests with 10,000 permutation samples. In all simulations, we treated the results obtained by the p^c_{perm} method as the underlying truth with which the performance of other methods was compared.

3.3.1 Type I error rate

Empirical type I error rates at the 0.05 significance level for 1000 replicates are summarized in the third column of Tables 2 and 3. The results clearly show that the type I error rates are significantly inflated for the χ^2_{2L} approximation under different simulation scenarios. The scaled chi-square approximation and the permutation procedure yield similar type I error rates which are close to the 0.05 nominal level. The two methods for correlation estimation have no significant effect on type I error rate.

Table 2. Empirical type I error rate and power for scenarios A and B under different sample sizes. The effects of β_j's are fixed at 0.1.

n	Methods	Model I	Model II	Model III
200	χ^2_{2L}	0.217	0.935	0.942
	$a\chi^2_g(\text{perm})$	0.051	0.787	0.785
	$a\chi^2_g(R^2)$	0.053	0.788	0.786
	Permutation	0.049	0.788	0.787
500	χ^2_{2L}	0.204	1.000	0.999
	$a\chi^2_g(\text{perm})$	0.052	0.994	0.991
	$a\chi^2_g(R^2)$	0.047	0.992	0.990
	Permutation	0.047	0.992	0.991

3.3.2 Power comparison

Table 2 summarizes the empirical power for scenarios A and B. The results obtained with the permutation method is considered as the underlying truth. It can be seen that the χ^2_{2L} approximation always gives the highest power (see column 3), which is due to its high false positive rate. The results produced by the scaled chi-square approximation are

very close to the permutation-based results, which indicates the good performance of the scaled chi-square approximation. No significant differences in power were observed for the two scaled chi-square approximation methods. However, the calculation with the $a\chi^2_g(R^2)$ method is much faster than the permutation-based $a\chi^2_g$(perm) method. The effect of sample size on power is clear: large sample size always gives large power, as we expected.

The results for scenarios C and D are summarized in Table 3. Similar trends as in Table 2 were observed. Again, the χ^2_{2L} approximation yields inflated false positive rates and is less attractive than the scaled chi-square approximation does. We also tried other correlations and found that negative or low positive correlations may reduce the overall power for given genetic effects. However, the overall trend as we observed in Tables 2 and 3 remains unchanged, when comparing the performance of different methods.

Table 3. Empirical type I error rate and power for scenarios C and D under different sample sizes. The effects of β_j's are fixed at 0.15.

n	Methods	Model I	Model IV	Model V
200	χ^2_{2L}	0.179	0.882	0.885
	$a\chi^2_g$(perm)	0.056	0.706	0.718
	$a\chi^2_g(R^2)$	0.053	0.703	0.709
	Permutation	0.054	0.704	0.714
500	χ^2_{2L}	0.189	0.998	0.996
	$a\chi^2_g$(perm)	0.057	0.986	0.989
	$a\chi^2_g(R^2)$	0.056	0.984	0.989
	Permutation	0.052	0.986	0.989

4. REAL DATA ANALYSIS

4.1 Dataset

We applied our method to a yeast data set generated for the purpose of eQTL mapping (Brem and Kruglyak 2005). The data were generated from 112 meiotic recombinant progenies of two yeast strains: BY4716 (BY; a laboratory strain) and RM11-1a (RM; a natural isolate) aimed at understanding the genetic architecture of gene expression. The data set contains expression profiles of 6216 gene expression traits and 2956 SNP markers. The readers are referred to Brem and Kruglyak (2005) for more details of the data set. The pathway information was retrieved from the R package: YEAST. There are 99 KEGG pathways in the package, but only 83 pathways were retrieved for follow-up analysis. The genotype profiles of neighboring markers tend to be highly correlated and some are even identical. With this information, markers were first merged to blocks (Sun 2007). Then missing genotypes were imputed based on available genotype information in each block. In cases where markers did not belong to any block, missing data were imputed by assuming a Bernoulli distribution with allele frequency estimated

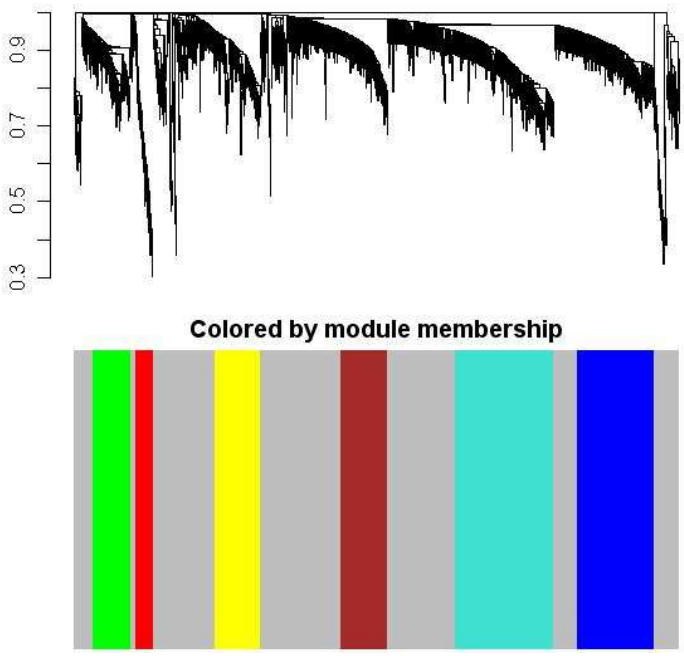

Figure 5. Weighted gene co-expression network with hierarchical clustering trees for the yeast gene expression data. See Zhang and Horvath (2005) for details of the algorithm.

based on available data for the corresponding marker. We focused our analysis on the pathway regulation of a network of genes as illustrated in Figure 1(E). We first built up gene expression networks using the gene expression traits. Then the method described in this work was applied to identify pathway regulations for each network.

4.2 Gene co-expression network

There are many ways to construct gene expression networks. We focused on gene co-expression networks following the method proposed by Zhang and Horvath (2005). Because of the computational burden, only the top 2001 connected genes out of the 4000 most varying genes were considered to build the co-expression networks. The average linkage hierarchical clustering method was applied to group genes with coherent expression profiles based on a topological overlap matrix (TOM) dissimilarity measure. In our study, we obtained six gene modules (Table 4). Figure 5 shows the six co-expression network modules. For a detailed description of the weighted gene co-expression network approach, the readers are referred to Zhang and Horvath (2005).

4.3 Network singular value decomposition

For each network, gene expression values were treated as multivariate responses and tested for association at each

Table 4. Information on gene co-expression networks

Modules	Blue	Brown	Green	Red	Turquoise	Yellow
# of genes	251	153	125	56	325	151
# of eigengenes	12	7	7	1	9	6

SNP marker locus. For the yeast data, there are two possible genotype categories at each locus. So a two sample Hotelling's T^2 test can be applied to test if mean responses are different for the two groups at each locus (Li et al. 2010). A gene co-expression network usually consists of many genes. In this dataset, most co-expression networks contain hundreds of genes. So the dimension of a network is greater than the sample size in most cases. Therefore it is infeasible to use Hotelling's T^2 test for expression profiles of all genes in a network. To reduce the dimension of a network, we applied the singular value decomposition (SVD) method. Because genes in a network are often highly correlated, using SVD could dramatically reduce the data dimensionality with only relatively few "eigengenes" capturing the total variation of a network. In this study, "eigengenes" that account for more than 85% of the total variation of a network of gene expression values were chosen as the response variable for further analysis.

Consider a gene expression network with N genes, all expression profiles can be represented by a matrix X with $N \times n$ dimension where n is the sample size. Each row of X represents the expression of one gene belonging to the network. The SVD of matrix X is given by

$$X = UDV^T,$$

where U is an $N \times K$ matrix; $D = diag\{d_1, d_2, \cdots, d_K\}$ is an $K \times K$ diagonal matrix, $d_1 \geq d_2 \geq \cdots \geq d_K$ are eigenvalues of X; and V^T is an $K \times n$ matrix with $K = min\{N, n\}$. Each row of matrix V^T represents a so-called "eigengene" of the original network. The proportion of "eigengenes" calculated by $v_k = d_k^2 / \sum_{k=1}^{K} d_k^2$ indicates the amount of total variation captured by the kth eigengene. Top J eigengenes will be remained for further analysis if the cumulative variation captured by the top J eigengenes is larger than 85%, i.e., $\sum_{j=1}^{J} v_j \geq 85\%$. The eigengenes are orthogonal to each other and are treated as a multivariate response to represent each co-expression network for further analysis.

4.4 Results by the scaled chi-square approximation

Hotelling's T^2 test was applied at each locus for gene expression networks with two or more eigengenes. For the red module with only one eigengene, a two-sided two sample t-test was applied. Individual p-values were then combined for each of the 83 genetic pathways to assess the significance by the scaled chi-square approximation. SNPs in different GPs may overlap which may cause dependence among GPs. The overlap issue was ignored in the current analysis and will be studied in future work. We also did the pathway enrichment analysis (PEA) proposed by Wang et al. (2007). The results are summarized in Table 5. Only GPs with p-values less than 0.001 were reported. The last three columns list the p-values for the combined statistic T using different methods to estimate the correlations plus those with the PEA analysis. The overlapped GPs with p-values less than 0.001 are highlighted with bold font. In many cases the enriched GPs identified with the two methods are very similar, except for the Blue module. In terms of the computation time, the combined p-value approach took much less time than the PEA analysis. For example, it took about 5 minutes to calculate the combined p-value with LD-based correlation approximation, while it took about 8 hours to run 1000 permutations for one network module with the PEA analysis.

We also tried the Fisher's χ^2_{2L} approximation assuming SNPs in a genetic pathway are independent. We found more significant pathways than with the scaled chi-square approximation (data not shown). As indicated by the simulation studies, the additional GPs identified are most likely false positives. From Table 5, we can see that pathways 78 (Pantothenate and CoA biosynthesis) and 20 (Valine, leucine and isoleucine biosynthesis) are responsible for several network expression changes. This implies the relative importance of these pathways in the regulation of yeast gene expressions.

In order to understand the biological significance of our findings, it is important that we first describe the origin of strains used in the original yeast crossing design. As mentioned earlier, the parental strains are derived from natural isolates. The first strain, BY4716, is a lab strain whose origin can be traced back to a natural isolate that was found growing on a rotting fig (Mortimer and Johnston 1986). However, this strain has had a long history of use as a laboratory model and has been selected for many properties that make it more amenable to experimentation (Barnett 2007). In addition, because it is derived from a haploid segregant of the original heterozygous, diploid natural isolate, and because it has been harbored in the relatively benign lab environment for many generations, several known loss-of-function alleles have been identified in this parental strain (Gu et al. 2005). Finally, all yeast strains used in experimental genetic crosses are altered to some degree. Most commonly these alterations include the generation of a null mutation for the HO endonuclease, the loss of which prevents mating type switching and allows for manipulation of ploidy and mating type (Burke et al. 2000). In addition, experimental yeast strains also harbor loss-of-function alleles for genes within amino acid biosynthetic pathways, so that nearly all lab strains are auxotrophic for some combination of amino acids (e.g., Uracil, Leucine, Lysine, Histine, Tryptophan, Methionine, Adenine) (Burke et al. 2000). Such auxotrophies provide a mechanism for phenotypic selection on yeast media that lacks specific amino acid supplements. Even though the second parental strain, a haploid derivative

Table 5. List of enriched genetic pathways (GPs) with the scaled chi-square approximation method and the gene set enrichment analysis. Only GPs with p-values ≤ 0.001 using either the p-value combined method or the PEA method are listed. The middle column is the list of GPs that are associated with the expression change of the corresponding co-expression networks given in the first column. GPs that show enrichment with both methods are highlighted with bold font.

Gene Network (# of genes)	P#	(PID)	Name of enriched GPs	$p_{a\chi^2_g}(R^2)$	$p_{a\chi^2_g}$(perm)	p_{PEA}
Blue	17	(03022)	Basal transcription factors	2.28e-03	1.75e-03	*< 0.001*
(251)	34	(04111)	Cell cycle - yeast	*7.55e-04*	3.03e-03	0.010
	78	(00770)	Pantothenate and CoA biosynthesis	*4.68e-04*	*7.69e-04*	0.011
Brown	10	(00500)	Starch and sucrose metabolism	8.89e-02	8.97e-02	*< 0.001*
(153)	**13**	(03020)	RNA polymerase	*2.53e-04*	*4.39e-04*	*< 0.001*
	17	(03022)	Basal transcription factors	*2.87e-04*	*3.69e-04*	*< 0.001*
	25	(00010)	Glycolysis / Gluconeogenesis	2.66e-02	3.05e-02	*< 0.001*
	32	(00920)	Sulfur metabolism	*7.11e-04*	1.12e-03	0.002
	34	(04111)	Cell cycle - yeast	*4.68e-05*	*2.81e-04*	*0.001*
	78	(00770)	Pantothenate and CoA biosynthesis	*3.97e-05*	*6.08e-05*	0.039
	83	(00220)	Urea cycle and metabolism of amino groups	*4.28e-04*	*6.41e-04*	*< 0.001*
	84	(00860)	Porphyrin and chlorophyll metabolism	*6.92e-04*	1.07e-03	*< 0.001*
Green(125)	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*3.50e-05*	*4.19e-05*	*< 0.001*
Red	**1**	(04010)	MAPK signaling pathway	*1.19e-04*	*1.06e-04*	*< 0.001*
(56)	10	(00500)	Starch and sucrose metabolism	1.23e-02	1.56e-02	*< 0.001*
	43	(00520)	Nucleotide sugars metabolism	*2.28e-05*	*3.53e-05*	*< 0.001*
	85	(00040)	Pentose and glucuronate interconversions	*3.04e-04*	*3.86e-04*	*0.001*
Turquoise	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*5.75e-07*	*3.45e-06*	*< 0.001*
(325)	**27**	(00650)	Butanoate metabolism	*6.40e-04*	1.30e-03	*< 0.001*
	78	(00770)	Pantothenate and CoA biosynthesis	*3.67e-05*	*1.43e-04*	0.002
Yellow	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*2.91e-39*	*1.05e-35*	*< 0.001*
(151)	**27**	(00650)	Butanoate metabolism	*1.92e-13*	*2.615e-13*	*< 0.001*
	74	(03010)	Ribosome	*2.99e-04*	*3.93e-04*	0.006
	78	(00770)	Pantothenate and CoA biosynthesis	*2.10e-19*	*6.41e-18*	*< 0.001*

P#=pathway number; PID=pathway ID.

of the natural vineyard isolate RM11-1a, was chosen to represent the prototrophic representative of a natural strain, it does carry loss-of-function alleles for HO endonuclease and auxotrophies for the Leucine and Uracil biosynthetic pathways (Brem and Krugylak 2002).

Strikingly, all of the pathways inferred to influence co-expressed gene groups can be traced to either the engineered or lab selected loss-of-function alleles segregating in the parental stains. For example, in Table 5, the Yellow gene co-expression module exhibited the highest statistical significance with respect to the functional categories that explain the observed variation. We did a GO term search and found that 43.7% of genes in this module are mapped to GO cellular amino acid and derivative metabolic process. This represents the highest percentage these genes can be mapped to the GO process category. Also 28.5% of genes are mapped to the GO transferase activity function category, which explains the enrichment of pathway 74 (Ribosome). KEGG genetic pathways 20 (Valine, leucine, and isoleucine biosynthesis), 27 (Butanoate metabolism), and 78 (Pantothenate and CoA biosynthesis) are all either directly requiring or downstream of the Lue2 (YCL018W) and Ilv6 (YCL009C) genes. These genes are both physically and functionally linked in that they are required for leucine and isoleucine biosynthesis and found with 13 kilobases of one another (roughly 3-5 centiMorgans) (Cherry et al. 1997). Because Leu2 is a complete knock-out, there were several markers all found within this locus, each strongly associated with a given pathway. Similarly, the Ilv6 gene, with only a single marker, is also strongly associated with all three of these KEGG genetic pathways. In addition, all or some combination of these genetic pathways are strongly associated with the Blue, Brown, Green, and Turquoise, gene co-expresssion networks, and in each case, the association is mediated by the same genetic markers. Hence a single engineered mutation that was known to be segregating in the parental cross explains most of the co-expressed genes in the Yellow module, and these same associations are found in the Blue, Brown, Green, and Turquoise

gene networks. All of these effects are likely mediated by a single loss-of-function at Leu2 with direct effect. In addition, the indirect effects of Leu2 on the regulation and activity of Ilv6 as well as the linkage of Ilv6 with Leu2 may also play an important role (Cullen et al. 1996; Cherry et al. 1997; Ronald and Akey 2007). Note that pathway 78 is enriched for the Blue, Brown and Turquoise network only by our approach, which indicates the better performance of our method against the PEA analysis in this study. Thus, this systems biology approach has allowed for the elucidation of many interacting gene networks and the genetic pathways through which they are most likely influenced. Importantly, these conceptual linkages derive from a clear biological reason, in this case an engineered mutation with pleiotropic effects.

In addition to the associations mediated via auxotrophic markers, the remaining genetic pathways can be broadly categorized in three groups: mitochondrial function (17 - Basal transcription factors; 13 - RNA polymerase), cell cycle (34 - Cell cycle), and cell signaling, filamentous / invasive growth, and mating (1 - MAPK signaling pathway). All of these effects are in pathways that can be traced to additional alleles of large effect that are known to have been segregating in the cross. Amn1 and Flo8 mutations in the lab strain were selected at some point in the past for reduced flocculation (clumpy growth due to cell-cell adhesion), and the 112 segregants differ in mating type at the MAT locus (Brem and Krugylak 2002; Mortimer and Johnston 1986). All of these selected and engineered alleles are known to be strongly involved in MAPK signaling. In fact, gene Ste20 (YHL007C) in the MAPK signaling pathway in this analysis shows the strongest single marker associations, and the gene is directly downstream of another well characterized QTL in previous studies, the Gpa1 gene (Wang and Dohlman 2004). Perhaps accidentally, the lab strain also is known to exhibit several phenotypes indicative of reduced mitochondrial function (Gaisne et al. 2000). While loss-of-function alleles were known to exist in the lab strain for the Hap1 (YLR256W) and Mkt1 (YNL085W) genes, a recent study mapping variation in mitochondrial function with these same data, identified three additional mitochondrial alleles of strong effect at Sal1 (YNL083W), Cat5 (YOR125C), and Mip1 (YOR330C), respectively (Dimitrov et al. 2009). In particular, Mip1 is part of the mitochondrial DNA polymerase and Hap1 is required for cytochrome function (Foury 1989; Pfeifer et al. 1989). Hence, the many genetic pathways related to mitochondrial function and localization (e.g., 92% genes in the Green module map to mitochondria via Gene Ontology) are likely a downstream pathway that was altered as a result of these known deficient alleles segregating in the cross. In this case, we suspect that given the importance of proper mitochondrial function in the wild, each of these alleles is due to relaxed selection in the lab environment (Ronald and Akey 2007).

Finally, the single largest effect size typically observed in studies utilizing data from this cross is at the Ira2 gene (YOL081W) (Ehrenreich et al. 2009). We observed very strong signals at this gene for all six co-expressed modules. The strongest one (p-value$< 10^{-14}$) corresponds to the Brown module. Even though this gene is not mapped to any KEGG pathways in this analysis, it is located upstream of the RAS/PKA signaling pathway and has strong downstream effects on nutrient signaling, cyclic AMP signaling, cell proliferation, and polymerase II activity (Broach 1991). The downstream effects of this polymorphism are apparent in the many genetic pathways related to nutrient metabolism, transcription, and cell cycle. Interestingly, this allele has not been traced to lab engineering or relaxed selection, but is more likely a naturally segregating difference that is derived in the vineyard isolate (Smith and Krugulak 2008).

In summary, our analysis has elucidated how a systems biology approach can identify the variation in genetic pathways that control co-expressed gene networks, and nearly all of the effects identified in this cross can be traced back to either engineered mutations or loss-of-function alleles that arose due to relaxed constraint in the benign lab environment.

5. DISCUSSION

The integration of gene expression analysis and genetic mapping, termed eQTL mapping, brings great promise in elucidating the genetic architecture of gene expression. Empirical studies have shown that eQTL mapping can shed new light into gene network prediction, provide additional biological insights into gene regulation, and facilitate functional gene identification (e.g., Bao et al. 2007; Chen et al. 2008; Schadt et al. 2008; Yang et al. 2009). Moreover, eQTL mapping results can provide additional directional information in gene regulatory network construction (Alberts et al. 2005; Keurentjes et al. 2007). With more biological data being generated at the sequence, transcriptional, proteomic and metabolic levels, together with the end-point phenotypic data such as a disease status, we are progressively approaching the era where various sources of data information can be integrated to gain novel biological insights from a systems biology perspective.

Our study is driven by the biological fact that genes function in networks or systems. Most biological phenomena occur through the expression of multiple genes which are potentially regulated by a cascade of genetic variants. Mootha et al. (2003) previously showed that focusing on expression data in terms of predefined pathways/networks (genetic features) can provide valuable insights into gene function not easily achievable by methods focused on individual genes. This inspired us to focus on features of genetic variants that belong to predefined pathways/networks in order to understand the genetic basis of gene regulation. Given the complexity of a genetic system, it is very unlikely that the function of a single variant will induce an overt identifiable or

physiologically meaningful expression change of a network of genes. Also features defined by groups of genes should be more robust to genetic variation. Thus, we proposed to incorporate pathway (e.g., KEGG pathway) information into an eQTL mapping framework to gain novel insights into pathway regulation of gene expression. By combining evidence of multiple signals in a genetic system, our method addresses the limitation of the traditional single marker–single trait analysis: 1) Without a single encompassing theme, results could be hard to interpret; 2) Moderate changes which were disregard in single marker analysis, may afford more insight into gene regulation mechanisms (Mootha et al. 2003).

As reviewed in the introduction section, there are many ways to combine evidences. It is commonly recognized that variants in a genetic system are often correlated. In this study we proposed to approximate the combined p-values of individual signals with a scaled chi-square approximation considering correlations among variants. Newton et al. (2007) proposed a random-set method in assessing gene-set enrichment by averaging gene scores. As discussed by the authors, among-gene dependence was not an issue in their enrichment analysis because factors that caused dependence were excluded from the calculation of a gene score. Instead of averaging, we proposed to combine signals. In addition, correlations among genetic variants preserved a structural relationship due to LD. Our simulation studies indicated that large false positive rates could be observed if correlations were not properly accounted for. We proposed two different methods for an estimation of the correlation information between the log-transformed p-values. The results indicate that using the LD information to approximate the correlation produces similar results as using permutation-based methods. Real data analysis also confirmed the result (Table 5). Thus, LD information could be directly applied in order to save computation time. It is also worth noting that depending on whether it is a one-sided or two-sided test, the relationship between the LD (R) and the correlation (ρ) could be different.

In the real data analysis, we focused our attention on gene expression networks as the response variables. We can also focus the responses on expression pathways extracted from public database such as those from KEGG database or from GO terms. Since only p-values are required, any sophisticated statistical tests can be applied. Even though the LD-based approximation for correlation of the log-transformed p-values may not be valid for a nonlinear model, the correlations can always be evaluated with the proposed permutation-based method. Depending on the interest of an investigator, our method provides a general strategy for regulation inference in a single gene or pathway level (e.g., Zhong et al. 2010). In addition, the method can also be extended to a (genome-wide) genetic association study to identify novel pathways underlying complex disease.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grant DMS-0707031 to YC and by an MSU QBI graduate fellowship to SL. We wish to thank the two anonymous referees for their helpful comments that greatly improved the manuscript.

REFERENCES

Alberts, R., Fu, J., Swertz, M.A., Lubbers, L.A., Albers, C.J., and Jansen, R.C. (2005). Combining microarrays and genetic analysis. *Brief. Bioinform.* **6**: 135-145.

Bao, L., Peirce, J.L., Zhou, M., Li, H., Goldowitz, D., Williams, R.W., Lu, L. and Cui, Y. (2007). An integrative genomics strategy for systematic characterization of genetic loci modulating phenotypes. *Hum. Mol. Genet.* **16**: 1381-1390.

Barnett, J.A. (2007). A history of research on yeasts 10: foundations of yeast genetics. *Yeast* **24**: 799-845.

Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc. Nat. Acad. Sci.* **102**: 15721577.

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**: 701-703.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752-725.

Broach, J.R. (1991). Ras-regulated signaling processes in Saccharomyces cerevisiae. *Curr Opin Genet Dev* **1**: 370-377.

Brown, M. (1975). A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**: 987992.

Burke, D., Dawson, D., and Stearns, T. (2000). Methods in Yeast Genetics A Cold Spring Harbor Laboratory Course Manual. CSHL Press.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., Leonardson, A., Castellini, L.W., Wang, S., Champy, M.F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T.A., Lusis, A.J., and Schadt, E.E. (2008). Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**: 429-435.

Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K., and Botstein, D. (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387**(6632 Suppl): 67-73.

Chun, H. and Keles, S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics* **182**: 79-90.

Conneely, K. N. and Boehnke, M. (2007). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *Amer. J. Hum. Gent.* **81**: 1158-2268.

Cullin, C., Baudin-Baillieu, A., Guillemet, E., and Ozier-Kalogeropoulos, O. (1996). Functional analysis of YCL09C: evidence for a role as the regulatory subunit of acetolactate synthase. *Yeast* **12**: 1511-1518.

Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., and Conklin, B.R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31**: 19-20.

De la Cruz, O., Wen, X., Ke, B., Song, M., and Nicolae, D.L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genet. Epidemiol.* **34**: 222-231.

Dimitrov, L.N., Brem, R.B., Kruglyak, L., and Gottschling, D.E. (2009). Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. *Genetics* **183**: 365-383.

Dudbridge, F. and Koeleman, B.P. (2003). Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol.* **25**: 360-366.

Ehrenreich, I.M., Gerke, J.P., and Kruglyak, L. (2009) Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross. *Cold Spring Harb Symp Quant Biol* **74**: 145-153.

Fisher, R.A. (1932). *Statistical methods for research workers*. London: Oliver and Boyd.

Foury, F.J. (1989). Cloning and sequencing of the nuclear gene MIP1 encoding the catalytic subunit of the yeast mitochondrial DNA polymerase. *Biol Chem* **264**: 20552-20560.

Gaisne, M., Bcam, A.M., Verdire, J., and Herbert, C.J. (1999). A 'natural' mutation in Saccharomyces cerevisiae strains derived from S288c affects the complex regulatory gene HAP1 (CYP1). *Curr Genet* **36**: 195-200.

Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**: 408-415.

Gu, Z., David, L., Petrov, D., Jones, T., Davis, R.W., and Steinmetz, L.M. (2005). Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae. *Proc Natl Acad Sci* **102**: 1092-1097.

Hess, A. and Iyer, H. (2007). Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *BMC Genomics* **8**: 96.

Hubner, N., Wallace, C.A, Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Mller, A., Cook, S.A., Kurtz, T.W., Whittaker, J., Pravenec, M., and Aitman, T.J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet.* **37**: 243253.

Jansen, R.C. and Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends Genet.* **17**: 388-391.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**: 277-280.

Kendziorski, C., and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome* **17**: 509-517.

Keurentjes, J.J., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci.* USA **104**: 1708-1713.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent P-values. *Stat. Prob. Lett.* **60**: 183-190.

Lee, E., Woo, J.H., Park, J.W., and Park, T. (2007) Finding pathway regulators: gene set approach using peak identification algorithms. *BMC Proc.* Suppl 1: S90.

Li, J. and Burmeister, M. (2005). Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet.* **14** Spec No.2: R163-169.

Li, S.Y., Lu, Q., and Cui, Y.H. (2010). A systems biology approach for identifying novel pathway regulators in eQTL mapping. *J. Biopharm. Stat.* **20**: 373400.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**: 267-73.

Mortimer, R.K. and Johnston, J.R. (1986). Genealogy of principal strains of the yeast genetic stock center. Genetics **13**: 35-43.

Newton, M.A., Quintana, F.A., den Boon, J.A., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat* 1: 85-106.

Mueller, M., Goel, A., Thimma, M., Dickens, N.J., Aitman, T.J., and Mangion, J. (2006). eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. *Bioinformatics* **22**: 509-511.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J.D., Jin, L., Amos, C.I., and Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* **18**: 111-117.

Petretto, E., Mangion, J., Pravanec, M., Hubner, N., and Aitman, T.J. (2006). Integrated gene expression profiling and linkage analysis in the rat. *Mamm. Genome* **17**: 480-489.

Pfeifer, K., Kim, K.S., Kogan, S., and Guarente, L. (1989). Functional dissection and sequence of yeast HAP1 activator. *Cell* **56**: 291-301.

Ronald, J. and Akey, J.M. (2007). The evolution of gene expression QTL in Saccharomyces cerevisiae. *PLoS One* **2**: e678.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J.D., Avila-Campillo, I., Kruger, M.J., Johnson, J.M., Rohl, C.A., van Nas, A., Mehrabian, M., Drake, T.A., Lusis, A.J., Smith, R.C., Guengerich, F.P., Strom, S.C., Schuetz, E., Rushmore, T.H., and Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**: e107.

Schadt, E.E., Monks, S.A., Drake, T.A. et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297-302.

Smith, E.N. and Kruglyak, L. (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol* **6**: e83.

Sun, W. (2007). Statistical strategies in eQTL studies. Ph.D. Thesis.

Thorn, C.F., Klein, T.E., and Altman, R.B. (2005). PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Methods Mol. Biol.* **311**: 179-191.

Wang, Y. and Dohlman, H.G. (2004). Pheromone signaling mechanisms in yeast: a prototypical sex machine. *Science* **306**: 1508-1509.

Wessel, J., Zapala, M.A., and Schork, N.J. (2007). Accommodating pathway information in expression quantitative trait locus analysis. *Genomics* **90**: 132-142.

Wu, C., Delano, D.L., Mitro, N., Su, S.V., Janes, J., McClurg, P., Batalov, S., Welch, G.L., Zhang, J., Orth, A.P., Walker, J.R., Glynne, R.J., Cooke, M.P., Takahashi, J.S., Shimomura, K., Kohsaka, A., Bass, J., Saez, E., Wiltshire, T., and Su, A.I. (2008). Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet.* **4**: e1000070.

Yang, X., Deignan, J.L., Qi, H., Zhu, J., Qian, S., Zhong, J., Torosyan, G., Majid, S., Falkard, B., Kleinhanz, R.R., Karlsson, J., Castellani, L.W., Mumick, S., Wang, K., Xie, T., Coon, M., Zhang, C., Estrada-Smith, D., Farber, C.R., Wang, S.S., van Nas, A., Ghazalpour, A., Zhang, B., Macneil, D.J., Lamb, J.R., Dipple, K.M., Reitman, M.L., Mehrabian, M., Lum, P.Y., Schadt, E.E., Lusis, A.J., and Drake, T.A. (2009). Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat. Genet.* **41**: 415-423.

Yu, K., Li, Q., Bergen, A.W., Pfeiffer, R.M., Rosenberg, P.S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* **33**: 700-709.

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

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* Vol.4, No.1, Article 17.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., and Weir, B.S. (2002). Truncated product mehtod for combining P-values. *Genet. Epidemiol.* **22**: 170-185.

Zhong, H., Yang, X., Kaplan, L.M., Molony, C., and Schadt, E.E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* **86**: 581-591.

Shaoyu Li
A401 Wells Hall
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824
E-mail address: lishaoyu@stt.msu.edu

Barry L. Williams
41 Giltner Hall
Department of Zoology and Microbiology and Molecular Genetics
Michigan State University
East Lansing, MI 48824
E-mail address: barryw@msu.edu

Yuehua Cui
A432 Wells Hall
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824
E-mail address: cui@stt.msu.edu