A systems biology approach for identifying novel pathway regulators in eQTL mapping

Shaoyu Li[1], Qing Lu[2] and Yuehua Cui[1*]

[1]*Department of Statistics & Probability,* [2]*Department of Epidemiology, Michigan State University, East Lansing, Michigan 48824*

Running head: Identify pathway regulators in eQTL mapping

Abstract

Expression quantitative trait loci (eQTL) mapping holds great promise in elucidating gene regulations and predicting gene networks associated with complex phenotypes. We propose a systems biology approach by incorporating prior pathway information into an eQTL mapping framework, to identify novel pathway regulators that mediate pathway expression changes. We model gene expressions in a pre-defined biological pathway as a multivariate response to test the joint variation changes among different genotype categories at a locus. The method is motivated and applied to a yeast dataset. Significant pathway regulators and regulation hotspots are detected. The proposed method provides a powerful tool for understanding gene regulations in a pathway level.

Key words: Expression quantitative trait loci, Hotelling's T^2 test, pathway enrichment analysis, Pathway regulator, Pathway regulation hotspot

*To whom correspondence should be addressed: Dr. Yuehua Cui, cui@stt.msu.edu

1 Introduction

Traditional quantitative trait loci (QTL) mapping has been focused on identifying genetic loci responsible for the phenotypic changes of a trait. Such studies are designed to detect linkage or association between genetic markers and the functional (causal) variants responsible for the phenotypic changes, and fail to disentangle the functional mechanisms of variants due to the regulation of other genes. In addition, the number and effect size for the detected alleles for most complex traits are very limited, leaving a large faction unaccounted for from a systems biology perspective.

Recent advances on microarray technology open an alternative front for multiple gene discoveries by studying thousands of gene expression profiles simultaneously under certain conditions or treatments. As an intermediate process that associates transcriptional profiles with an organism's trait variation, analysis of gene expression holds great promise to infer genetic regulatory changes accompanying a disease trait, and serves as an alternative to identify novel relationships among genes. A number of studies have shown that gene expressions are inheritable traits, thus can be used for genetic mapping (e.g., Brem et al., 2002; Cheung et al. 2003; Schadt et al. 2003). The two endeavors, genetic mapping and gene expression analysis, were recently merged together through a procedure called expression QTL (eQTL) mapping in which each gene expression is considered as one trait for QTL identification (Schadt et al., 2003).

Most current eQTL mapping studies treat each gene expression as one single trait. The so called single trait analysis may not be powerful enough to identify genetic variants responsible for gene expression changes, given that genes function in networks. Wessel et al. (2007) found in their eQTL mapping study that many SNPs are responsible for the expression change of genes belonging to a certain pathway. It is commonly recognized that genes in a biological pathway, e.g. metabolic pathway, gene regulation pathway or signal transduction pathway, "cooperate" with each other and function as a team to fulfill their designated tasks. Differential expression of one gene, especially those that play key roles in the pathway, would influence expression levels of other genes in the same pathway. Thus, a signal perturbation of a particular gene in a pathway would induce a cascade of biochemical events that affects all, or many of the other genes belonging to the same network or pathway. Take this functional mechanism of a pathway into account, the currently broad claims of cis- or trans-regulation detected with the single trait analysis might not be sufficient and efficient enough to capture the relationship between genetic variations and gene expressions.

Mootha et al. (2003) have previously showed that focusing on expression data in terms of predefined pathways can provide valuable insights not easily achievable by methods focused on individual genes. Many scientists are thus interested in identifying which genetic variant mediates the expression change of a pathway. The identified regulator, termed pathway regulator, provides additional information about the function of gene regulation from a systems biology perspective. A number of eQTL studies have taken pathway information into account (e.g. Lee et al., 2007; Wu et al., 2008). Most of these studies follow a two-stage procedure: do a single trait analysis in the first stage and then perform a gene set enrichment analysis (GSEA) to test if an expression pathway is enriched at a particular locus. The two-stage approach obviously does not take the expression correlation information into account; even genes in a pathway are commonly correlated. Moreover, the accuracy of the second stage enrichment analysis depends heavily on the results of the first stage. When genes function jointly but with small marginal effects, this

approach may fail to identify important pathway regulators. The relative merits of multivariate analysis over the univariate analysis is clearly demonstrated in the example given in Johnson and Wichern (2007, example 6.16 and 6.17 on page 333-335). Another disadvantage of the two-stage analysis is the multiplicity issue. With thousands of gene expression profiles, single trait analysis needs to adjust for the large number of tests when declaring significance. This may lead to a low power in identifying genes with small marginal effects, which again affects the power of the second stage enrichment analysis.

Considering the limitation of the current single-trait-based analysis and motivated by real biological phenomenon, we propose to identify common pathway regulators by treating gene expressions belonging to a common pathway as a multivariate response, and focus our interests in identifying pathway regulators that mediate the expression changes of a particular biological pathway or process. More importantly, when multiple gene expressions are jointly considered, the multiple testing burden in a single trait analysis is potentially reduced, hence leading to increased power. For the illustrations in the paper we restrict ourselves to one yeast dataset (Brem and Kruglyak, 2005). Our analysis indicates there are potential pathway regulators in regulating pathway gene expressions. Significant pathway regulator hotspots are identified. We also performed an enrichment analysis to test which genetic pathway is enriched in regulating the expression change of a certain expression pathway, and found significantly enriched genetic pathways in regulating other pathway gene expressions.

2 Methods

2.1 eQTL dataset

The yeast dataset was generated from 112 meiotic recombinant progeny of two yeast strains: BY4716 (BY; a laboratory strain) and RM11-1a (RM; a natural isolate), aimed to understanding the genetic architecture of gene expressions. The dataset contains expression profiles of 6216 gene expression traits and 2956 SNP marker genotype profiles. For details about the dataset, see Brem and Kruglyak (2005).

In the yeast genotype profiles, genotypes of neighboring markers tend to be very similar and some are even identical. For those SNP markers showing high correlations, we follow the strategy proposed by Sun (Sun, 2007) to construct marker blocks, in order to remove redundancy and reduce the genotype dimension. Specifically, we

i) *Merge markers into marker blocks*: Define $u = (u_1, u_2, \cdots, u_n)^T$ and $v = (v_1, v_2, \cdots, v_n)^T$ as vectors of two SNP genotype profiles over n individuals. Each SNP is coded as 0 or 1 depending on whether it is inherited from the BY or the RM strain. The Manhattan distance between the two SNP genotype vectors is defined as

$$MD = \sum_{i=1}^{n} |u_i - v_i|$$

The value of *MD* indicates the degrees of overlap between the two SNP markers. A small value indicates much overlap between the two markers. We include a SNP marker into a marker block if the Manhattan distance between a marker and any markers in its neighborhood is less than a predefined value d. In our analysis, we set $d = 1$. Other values like 1.25, 1.5 could also be used, depending on how strict constrain you want to put on the marker

similarity. If either u_i or v_i is missing for any individual i, the term $|u_i - v_i|$ is excluded from the summation, and the *MD* measure is adjusted by multiplying a factor $n/(n\text{-}m)$ (Sun, 2007), where m is the total number of terms been excluded. With d=1, we ended up with 1168 marker blocks.

ii) *Define genotype profiles for each marker block*: We first find consensus for each marker block and then dichotomize it. An individual genotype is set to 0 or 1 if at least 75% of the markers in a block equals to 0 or 1 for that individual. Otherwise it is set as missing. Individuals with missing genotypes will be eliminated.

Quite often no markers belong to any blocks, and some marker blocks may consist of multiple markers. We interchange the two words "marker" and "marker block" in the following presentation, but both terms represent one set of marker.

2.2 Genome-wide pathway regulator identification

We focused our analysis on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. We extracted 99 pathways from the R package: YEAST. Let $Y_i = \{y_{i1}, \cdots, y_{ip}\}^T$ be a vector of gene expressions in a pre-defined pathway for the ith subject, where p is the size of the pathway. We assume that the 1168 SNP marker blocks are causal or closely linked to the genetic variants responsible for the gene expression changes. For each SNP marker, there are two genotype categories, with each one corresponding to one multivariate expression profile. To test the differential expression pattern between different genotypes at a locus, a Hotelling's T^2 test can be applied which has the form

$$T^2 = (\overline{Y}_0 - \overline{Y}_1)^T [(\frac{1}{n_0} + \frac{1}{n_1}) S_{pooled}]^{-1} (\overline{Y}_0 - \overline{Y}_1) \tag{1}$$

Where $\overline{Y}_j = \sum_{i=1}^{n_j} Y_i$ and n_j are the sample mean expression vector and sample size for genotype coded as j (j = 0, 1). Assuming equal variance for expression values in the two genotype categories, pooled variance estimation S_{pooled} can be used in defining the T^2 statistics. The T^2 test is performed for all 99 KEGG pathways at every marker across the whole genome. Noted that the T^2 test is a two sided test. Even though genes in the same pathway may be up- or down-regulated, the test is still valid since we are interested in testing the mean vector difference between two genotype groups.

Theoretically the T^2 statistic follows a scaled F distribution (Johnson and Wichern, 2007). To control the family-wise error rate across the whole genome, we perform a permutation test to determine the genome-wide cutoff. When doing permutations, each row vector of gene expression is considered as one observation to retain the gene correlation information within a pathway. Then we fix the genotype information and randomly sample expression vectors without replacement. This random reshuffling procedure disturbs the relationship between gene expressions and genotypes. One thousand permutations are conducted to generate a null distribution for the T^2 statistic. For each permutation, T^2 values for all marker blocks are calculated, and the maximum T^2 value is recorded. The 1000 maximum T^2 values represent the genome-wide null distribution of the T^2 statistic in which the 95th percentile is considered as the genome-wide cutoff. A SNP marker block is considered as a pathway regulator if the observed T^2 value is greater than the cutoff value.

The T^2 test is performed when the number of genes in a pathway is less than the sample size. However, in real application some pathways may contain large number of genes ($p > n$). Given the small sample size (total of 112) in the yeast dataset, this dose happen (e.g., pathway '04111' and '03010'). When this happens, instead of using the T^2 statistic, we can apply the F statistic proposed by Zapala and Schork (2006). Consider a multivariate regression model

$$Y = X\beta + \varepsilon \tag{2}$$

where Y is a multivariate response (e.g., gene expressions in a pathway), X is the design matrix for SNP genotypes. When only one SNP marker is considered, X is an $n \times 2$ matrix with ones in the first column and numerical genotype coding in the second column. The F statistic proposed by Zapala and Schork (2006) has the form

$$F = \frac{tr(HGH)}{tr[(I - H)G(I - H)]}$$

where H is the hat matrix of the multivariate regression model in (2), and

$$G = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')A(I - \frac{1}{n}\mathbf{1}\mathbf{1}')$$

Matrix $A = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$, which is a so called distance matrix (or dissimilarity matrix) that measures distance (or dissimilarity) between expression levels of genes in a pathway; $\mathbf{1}$ is a column vector of ones and I is an identity matrix. An easy way to form the distance matrix is to use the correlation matrix and transform them with simple transformation technique, i.e. $d_{ij} = \sqrt{2(1 - r_{ij})}$ where r_{ij} is the correlation between genes i and j (Zapala and Schork, 2006).

The F statistic is especially useful when the number of parameters p is larger than the sample size n (Zapala and Schork, 2006). However, it is not trivial to find the theoretical distribution of the F statistic. Here, we still conduct a permutation procedure to assess the statistical significance. When $p = 1$, the F and the T^2 statistics are identical if the distance matrix is computed by the standard Euclidean distance measure. For pathways with small number of genes, results obtained with the two statistics are also very consistent. However, the F method is more time demanding due to large matrix operation for pathways with large number of genes. Thus, we only apply this method to pathways with $p > n$.

2.3 Pathway regulation hotspot detection

An eQTL hotspot is defined as a genetic region where a large number of gene expressions are mapped to than random (Morley *et al.*, 2004; Breitling *et al.*, 2008). In traditional eQTL hotspot detection, genomic regions are generally defined as "bins" with each "bin" covering a genomic interval in a length of, say 5Mb (in humans) (Morley *et al.*, 2004). Similar as the regular eQTL hotspot detection, we can also identify pathway regulation hotspot. Let $N_l, (l = 1, \cdots, L(=1168))$ be the number of pathways which are significantly mapped to marker block l. Let $N = \sum_{l=1}^{L} N_l$ be the total number of pathways significantly mapped to the whole genome. Then a Poisson distribution can be assumed for each N_l with the mean parameter λ estimated by the empirical mean N / L. Consider each marker

block as one potential hotspot, the probability of observing N_l or more significant pathways mapped to a marker block can be considered as the hotspot p-value, denoted as p_l. Take the Bonferroni correction at the 0.05 genome-wide significant level, a marker block is considered as a pathway regulation hotspot if $p_l < 0.05/L$. Alternatively, we can combine neighborhood marker blocks as one synthetic block with a pre-defined length, e.g., 20kb length. Then the total genome can be divided into K ($< L$) segments. Following the same procedure described above, pathway regulation hotspots can also be tested.

Relaxing the Poisson assumption, we can also use a nonparametric permutation procedure to identify regulation hotspots. Let $Q = (q_{ij})$ be a matrix which contains the mapping results, where $q_{ij} = 1$ if pathway $i, (i = 1, 2, \cdots, 99)$ is significantly mapped to locus $j, (j = 1, \cdots, 1168)$ and $q_{ij} = 0$ otherwise. We randomly permute the positions for 1's for each row of matrix Q and generate 1000 permuted matrices $Q_1^*, Q_2^*, \cdots, Q_{1000}^*$ while keeping the row sums of all these Q_p^*'s the same as the row sum of original observed matrix Q, i.e.

$$\sum_{j=1}^{1168} q_{p,ij}^* = \sum_{j=1}^{1168} q_{ij}, i = 1, 2, \cdots, 99, p = 1, 2, \cdots, 1000$$

The distribution of column sums for each permuted matrix is recorded. A locus is declared as a regulation hotspot if the observed count at that locus is larger than the 95th percentile of the permuted distribution.

2.4 Genetic pathway enrichment analysis

In a recent genome-wide association study for identifying disease risk variants, Wang *et al.* (2007) proposed a pathway-based association study to map genetic pathways (GPs) involving multiple genetic variants functioning together to give rise to a disease phenotype. In reality, we expect certain genetic pathways be enriched in responsible for the expression change of an expression pathway. By genetic pathway we mean SNP variants that belong to a common pathway. Here we use GP to denote a genetic pathway and use EP to denote an expression pathway. The purpose of this analysis is to identify which GP is enriched in mediating the expression change of an EP. For an enriched GP corresponding to an EP, we anticipate the expression variation of the EP can be explained by the joint function of SNPs in the GP.

From the genome-wide analysis, we can obtain a list of significant pathway regulators (markers) corresponding to an EP. Total 1465 unique genes (including annotated and non-annotated) are extracted from the whole genome. GPs are then grouped according to the KEGG pathway information. We call a gene is significant if there is at least one marker in this gene is significant. It is possible that several markers in a gene are significant. Similarly, there are total of 99 GPs retrieved from the KEGG database. Fixing an EP, we test which GP is enriched to explain the expression variation for that EP. Let n_S be the total number of genes that are significantly associated with an EP. Let n_G be the number of genes that belong to a GP, among which S are significantly associated with an EP. Then we can formulate a 2×2 table shown in Table 1. The Fisher's exact test

can be applied to calculate the enrichment p-value which is then compared with a significance level α. We use a less conservative α value, i.e., α = 0.01 to declare GP enrichment.

Table 1: A simple layout for testing genetic pathway enrichment

	No. of genes in a GP	No. of genes not in a GP	Total
No. of Sig. genes	S	$n_S + S$	n_S
No. of non-sig. genes	$n_G - S$	$K - n_G - n_S + S$	$K - n_S$
Total	n_G	$K - n_G$	K

K (= 1465) is the total number of unique genes covering the marker blocks across the genome.

3 Results

3.1 Pathway regulators

There are totally 99 pathways retrieved from the Yeast package in R for this dataset. The names of the 99 pathways are listed in the supplement table (Suppl. - Table 1). At each marker block, a T^2 or F statistic is calculated for each gene expression pathway, depending on the size of the pathway. We illustrate the idea with one pathway: MAPK signaling pathway. Figure 1 shows the T^2 profile plot for the pathway across the 16 yeast chromosome. The horizontal dash-dotted line in the plot indicates the 5% genome-wide threshold by permutation tests. Genomic positions where the T^2 peaks passing the threshold are considered as potential pathway regulators. For this pathway, we identified several pathway regulators on chromosome 2, 3, 5, 8, 14, 15 and 16. A full plot of the T^2 or F profiles for all the 99 pathways are listed in the supplement file (Suppl.- Figure 1). All the T^2 (or F) values as well as the genome-wide cutoffs are log transformed with base 10. In Suppl.-Figure 1, the minimum value for the vertical axis is truncated to the mean of the log10 transformed values. The cutoff values for each pathway are also labeled.

As indicated in the supplemental figure (Suppl. - Figure 1), we can see consistent strong association signals on chromosome 3, 14 and 15 which indicates that there are important pathway regulators located on these three chromosomes. Since a large number of EPs are regulated by these regulators, they are potentially "master" pathway regulators. For example, SNP marker YCL009C (belongs to gene ILV6) on chromosome 3 regulates 39 EPs and its neighborhood genes (e.g., genes LEU2 and BUD3) also regulate large number of EPs.

Figure 2 shows how many regulators each EP has. All the 99 EPs are plotted in the horizontal axis and the vertical axis indicates the number of regulators each pathway has. We can clearly see that the expressions of some pathways are affected by many genetic variants. For example, pathway 62 (Pyruvate metabolism pathway) has 98 regulators. Some pathways are not regulated by any variants, for instance, pathway 60 (ABC transporters - General) and 90 (Two-component system - Organism-specific). Note that many markers are highly correlated in this yeast dataset. Even though we merged some markers with large proportion of overlaps, we still expect large number of markers to be highly correlated with neighborhood markers. Thus, Fig. 2 only gives us a rough idea of how each pathway expressions are affected by many regulators. Whenever there is a causal regulator presented in a genomic region, due to strong linkage disequilibrium (LD) between neighborhood markers, its

neighborhood markers might also show strong association signals. Thus, the true regulators for each EP might be smaller than the reported numbers.
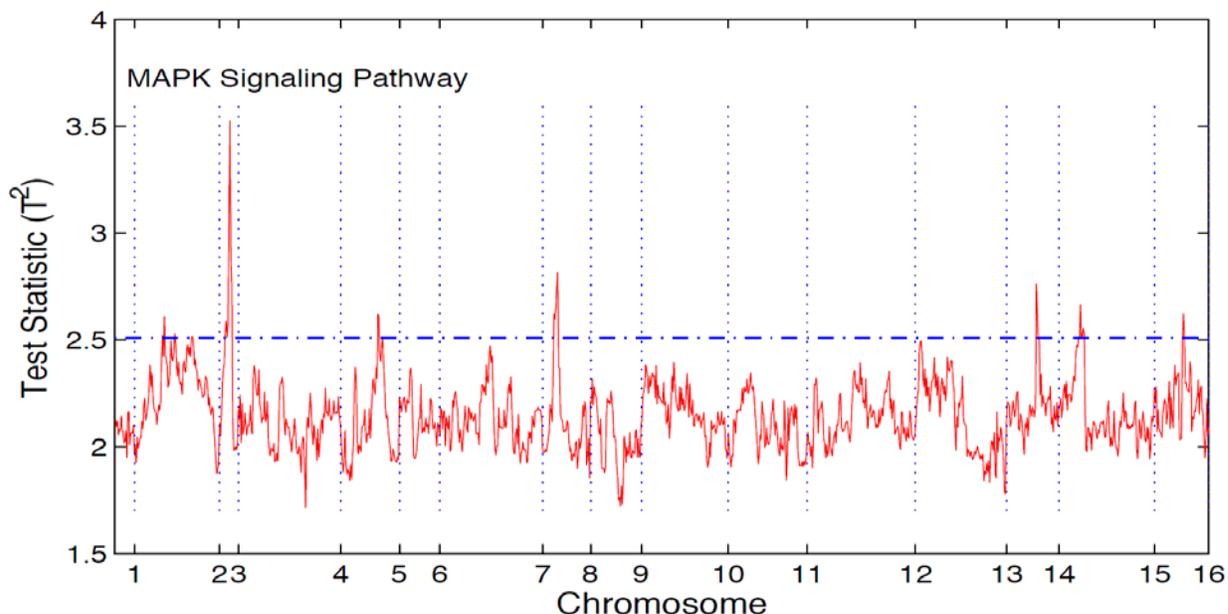

Figure 1: The T^2 profile plot across the entire yeast genome (16 chromosomes). The dash-dotted horizontal line is the 5% genome-wide permutation cutoff. The vertical dotted lines separate different chromosome regions. The peaks of the T^2 profiles that pass the cutoff correspond to potential pathway regulation loci (e.g. on chromosome 2, 3, 5, 8, 14, 15 and 16). Both the cutoff and the T^2 values are $\log_{10}$ transformed.

3.2 Pathway regulation hotspots

In eQTL mapping study, people are often interested in knowing which genomic region or interval plays an important role in regulating gene expressions. The so identified regions or intervals are called eQTL hotspots (Morley *et al.*, 2004). Since we merged some markers to form marker blocks, we simply treat each block as one potential pathway regulation hotspot and assess its significance. We count the number of EPs regulated by each marker block across the genome. The average number of association for one marker block is $\hat{\lambda} = 2.52$, and none of the marker block was expected to contain association with more than 10 pathways by chance at the 5% genome-wide significant level after Bonferroni correction (0.05/1168). We detect a total number of 76 pathway regulation hotspots. As shown in Figure 3 the distribution of the identified hotspots. The horizontal dash-dotted line indicates the threshold calculated from the Poisson model. The vertical bars indicate the number of pathways regulated by each marker block. Significant pathways at the hotspots are indicated by red color and all other significant pathways are indicated by cyan color. We identified several pathway regulation hotspot groups located on chromosome 2, 3, 5, 10, 12, 13, 14 and 15. Chromosome 5 and 15 show two distantly located hotspots.

It is interesting to note that most of the hotspots are clustered together on the genome. Some clusters have narrow band (e.g., the ones on chromosome 5, 10, 14 and 15), and some have wide band
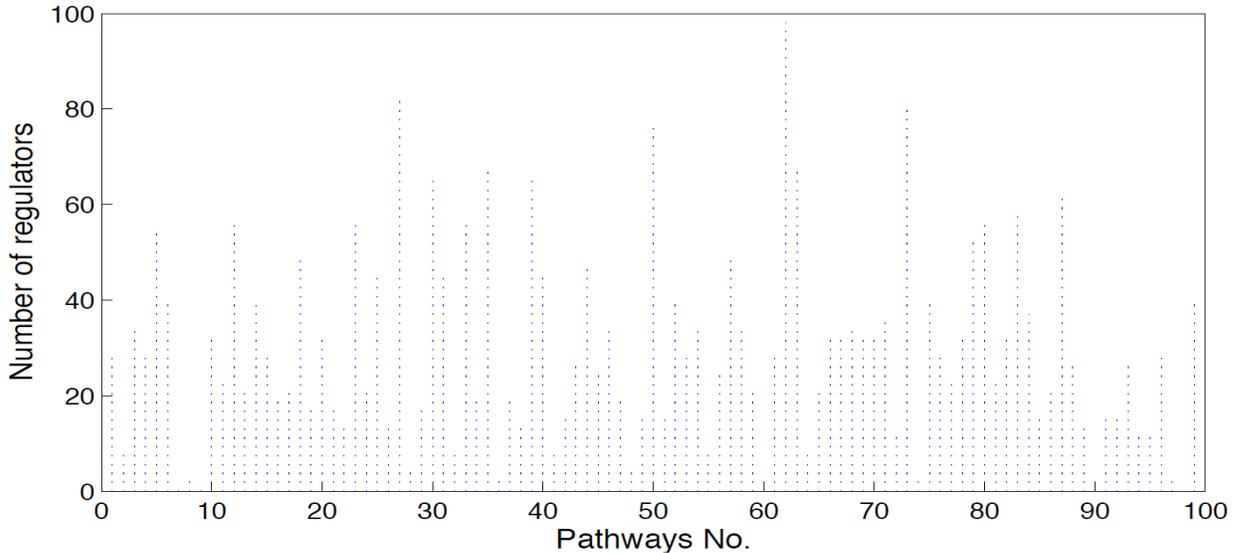
Figure 2: Number of regulators for each expression pathway. The horizontal axis denotes the 99 KEGG pathways and the vertical axis denotes the number of marker blocks that are significantly associated with each expression pathway.

(e.g., the ones on chromosome 2, 12 and 13). As we noted from the marker data, there are strong correlations between markers for this yeast dataset. Thus, this kind of pattern is expected. If we increase the hotspot interval size, the hotspot bands would become narrower with sharper peaks.

We also applied the permutation method to detect regulation hotspot. When using the permutation method, the cutoff is changed to 11. The horizontal dashed line in Fig. 3 indicates the permutation cutoff. Loci with more than 11 associated pathways were identified as hotspots. With increased threshold, the number of regulation hotspots reduced to 67. Several hotspots including the ones on chromosome 5 and 10 are no longer significant with the new threshold value. Overall, the two methods for regulation hotspot identification give quite similar results. A detailed list of the hotspot regulation is given in the supplemental file (see Suppl. Table 2).

In a recent study of genetic basis for small-molecule drug response, Perlstein *et al.* (2007) detected eight QTL hotspots located on chromosome 1, 3, 12, 13, 14 and 15 with the same yeast marker data. Five of those (on chromosome 3, 12, 13 and 14) overlap with the hotspots we identified. This information indicates the relative importance of these four genomic regions in regulating gene expressions as well as drug response. It is possible that the variation in drug response is due to the variation in pathway expressions which are directly related to hotspots regulation. Models can be developed to test the causal relationship among the three sets of data including genetic, gene expression and clinical phenotype (see e.g., Schadt *et al.*, 2005). We will consider this in our future work.

3.3 Genetic pathway enrichment

In addition to identify pathway regulation hotspots, we also perform a functional enrichment analysis using the Fisher's exact test to assess if a GP is enriched in regulating the expression of an EP. An enrichment p-value is computed to reflect the degree in which a given GP is over-represented.
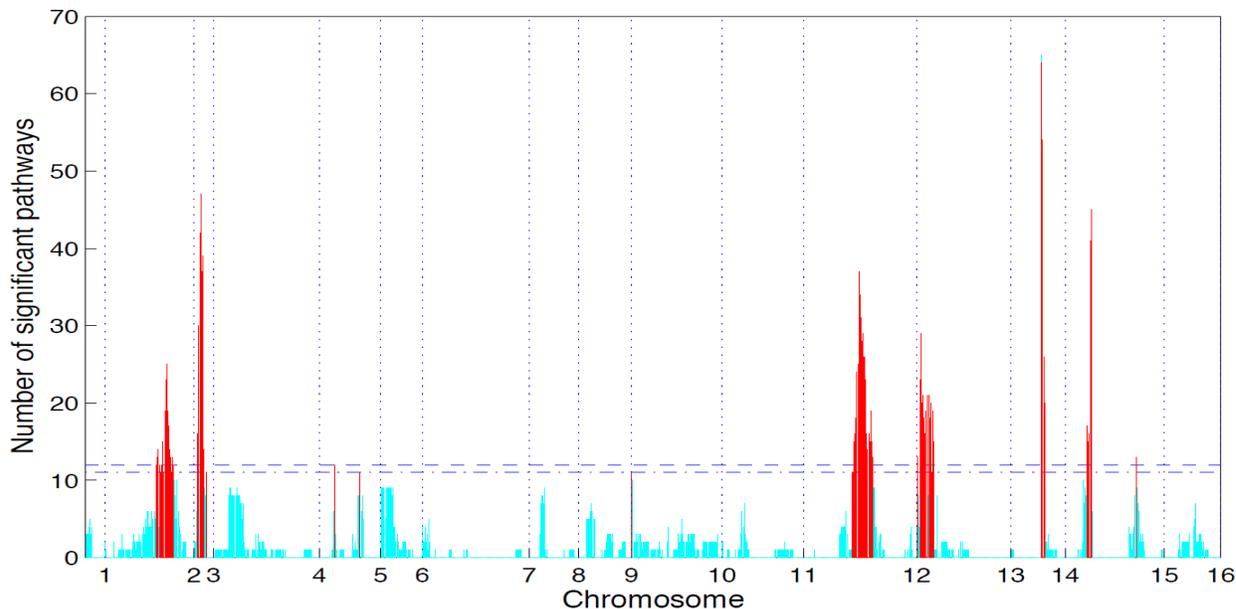
Figure 3: The pathway regulator hotspot. The dash-dotted and the dashed lines indicate the threshold calculated using the Poisson distribution and the permutation method, respectively. Dark black bars indicate the regulation hotspots. Vertical axis indicates the number of regulated EPs.

The results are tabulated in Table 2. A heatmap of the pathway enrichment analysis is given in the supplemental figure (Suppl. - Figure 2). To make the table consistent with the supplemental figure (Suppl. - Figure 2), we also list the pathway number (denoted as #) in addition to the pathway identification number (PID). The left column shows the enriched GPs which are responsible for the expression change of the corresponding EPs in the left column. All enriched GPs are claimed at the 1% significance level.

Clearly pathway 15, 20 and 74 are relatively important in regulating the expression of other pathways, since each one is enriched for a large number of EPs. It is hypothesized that the signal perturbation of these pathways may have pleiotropic consequences on multiple downstream pathways. Particularly for pathway 20, it may act as potential "master" pathway regulators as it regulates 25 pathway expressions. Also noted that most enriched GPs are metabolism or biosynthesis related pathways, which may indicate that these pathways might play key roles in the yeast genome.

In testing GP enrichment, we found that some GPs are enriched in regulating its own gene expressions. We define these GPs who regulate their own gene expressions as *cis*-pathway regulators. The highlighted bold-font pathways in Table 2 are those that show strong *cis*-regulation effects. These six GPs are pathway 13, 15, 20, 27, 43 and 44. All the others show *trans*-regulation effects. Note that the enriched GPs are claimed at the 0.01 significance level. If we lower the significance level to 0.001, all the *cis*-regulation pathways are gone, indicating that the *cis*-regulation effect is actually weaker than the *trans*-regulation effect in this application.

In a closer look at the enriched GPs, we found that pathway 20 contains two SNP markers (YCL009C in gene ILV6 and YCL018W in gene LEU2) that are located on the hotspot on chromosome 3. LEU2, beta-isopropylmalate dehydrogenase, plays an important role in catalyzing the third step in the leucine biosynthesis pathway. Pathway 78 also contains two SNP markers (YBR176W in gene ECM31 and YCL009C in gene ILV6). In checking the KEGG pathway, we

10

found that ILV6 is on the most upstream in pathway 78. Thus this gene may play a key role in affecting the downstream gene functions and in turn affecting many other pathway expressions.

4 Discussion

Understanding the genetic architecture of complex traits is one of the major challenges in modern biology. In a series of recent advances, many efforts have been focused on mapping genetic regions, called QTLs, in responsible for the phenotypic variation of a complex trait. Due to limited mapping resolution and with other non-genetic factors contributing to the phenotypic variation, this process has not been very successful in real applications, leaving only a few successful cases being reported in literature (e.g., Frary *et al.*, 2000; Li *et al.*, 2006). Recent advances in microarray technology allows us to measure the transcription abundance of many organisms and hence open another framework in understanding the genetic basis of gene expression, aimed to understanding the regulation of a genetic system. The initiation of the eQTL mapping with combined genetic mapping and gene expression analysis brings new prospects in understanding the complex process of gene regulation toward the ultimate goal of improving trait quality and disease prevention (Cookson *et al.*, 2009).

Based on the biological assumption that genes function in networks, dissect the genetic architecture of gene regulation from a systems biology perspective should provide more insights regarding the function of a biological system. In this article, we made an attempt to study gene regulations by combining gene expression and genetic polymorphism data together and proposed a pathway-based systems biology approach that aims to identifying genetic variants that regulate pathway gene expressions. We proposed to do the analysis by considering gene expressions in a pre-defined pathway as a multivariate response. Since genes in the same biological pathway tend to have similar expression pattern, looking at a bunch of expression levels in a pathway as our unit phenotype will give us more information about the differential expression pattern about this pathway, and thereby will give us more power to steadily detect the association of a genetic variation with the expression changes of a pathway. We focused our application to a real dataset in yeast and identified significant regulation patterns across the 16 yeast chromosomes. The detected pathway regulators tend to cluster together on the genome which might be due to the strong correlations among SNP markers on the genome.

Strong pathway regulation hotspots were identified in this study. Most of the hotspots overlap with the ones tested with single trait analysis (Brem and Kruglyak, 2005). Perlstein *et al.* (2007) recently applied the same yeast data to study individual genetic differences in response to small-molecule drugs and identified eight hotspots in response to multiple com-pounds. Their hotspots overlap with most of the pathway regulation hotspots identified in our study except the one on chromosome 1. This information indicates that the same polymorphisms may affect both gene expression and compound response. The genetic enrichment test proposed in this work can be applied to their study to understand which genetic pathways are involved in drug response. Noted in our analysis, each hotspot contain either a single pleiotropic polymorphism or several closely linked polymorphisms (marker block) affecting the response to multiple pathway expressions. If we group

Table 2: Genetic pathway enrichment analysis results. The right column indicates enriched genetic pathways (GPs) that are responsible for the expression change of the corresponding expression pathways (EPs) in the right column, at the 0.01 significant level. Pathways highlighted with bold faces indicate *cis*-pathway regulation.

#	(PID) Enriched Genetic Pathway	#	(PID) Expression Pathway
1	(04010) MAPK signaling pathway	47	(00480) Glutathione metabolism
13	(03020) RNA polymerase	**13**	**(03020) RNA polymerase**
15	(00051) Fructose and mannose metabolism	**15**	**(00051) Fructose and mannose metabolism**
		16	(00052) Galactose metabolism
		43	(00520) Nucleotide sugars metabolism
		70	(00625) Tetrachloroethene degradation
16	(00052) Galactose metabolism	15	(00051) Fructose and mannose metabolism
		43	(00520) Nucleotide sugars metabolism
17	(03022) Basal transcription factors	83	(00220) Urea cycle and metabolism of amino groups
20	(00290) Valine, leucine and isoleucine biosynthesis	1	(04010) MAPK signaling pathway
		4	(00910) Nitrogen metabolism
		6	(00410) beta-Alanine metabolism
		18	(00053) Ascorbate and aldarate metabolism
		20	**(00290) Valine, leucine and isoleucine biosynthesis**
		25	(00010) Glycolysis / Gluconeogenesis
		26	(00330) Arginine and proline metabolism
		32	(00920) Sulfur metabolism
		38	(00563) GPI-anchor biosynthesis
		45	(00340) Histidine metabolism
		49	(00750) Vitamin B6 metabolism
		52	(00251) Glutamate metabolism
		54	(00252) Alanine and aspartate metabolism
		58	(00670) One carbon pool by folate
		61	(00300) Lysine biosynthesis
		66	(00260) Glycine, serine and threonine metabolism
		73	(00310) Lysine degradation
		75	(00630) Glyoxylate and dicarboxylate metabolism
		77	(00450) Selenoamino acid metabolism
		78	(00770) Pantothenate and CoA biosynthesis
		86	(00360) Phenylalanine metabolism
		88	(00680) Methane metabolism
		92	(00401) Novobiocin biosynthesis
		96	(00272) Cysteine metabolism
		99	(00903) Limonene and pinene degradation
27	(00650) Butanoate metabolism	**27**	**(00650) Butanoate metabolism**
35	(00561) Glycerolipid metabolism	51	(00521) Streptomycin biosynthesis
		68	(00530) Aminosugars metabolism
39	(00513) High-mannose type N-glycan biosynthesis	87	(01030) Glycan structures - biosynthesis 1
43	(00520) Nucleotide sugars metabolism	**43**	**(00520) Nucleotide sugars metabolism**
44	(00020) Citrate cycle (TCA cycle)	**44**	**(00020) Citrate cycle (TCA cycle)**
46	(00980) Metabolism of xenobiotics by cytochrome P450	64	(00440) Aminophosphonate metabolism
68	(00530) Aminosugars metabolism	53	(00072) Synthesis and degradation of ketone bodies
74	(03010) Ribosome	45	(00340) Histidine metabolism
		86	(00360) Phenylalanine metabolism
		92	(00401) Novobiocin biosynthesis
78	(00770) Pantothenate and CoA biosynthesis	86	(00360) Phenylalanine metabolism
		92	(00401) Novobiocin biosynthesis

PID=Pathway ID

the genomic regions as intervals with 20kb length as previous work did (see Brem and Kruglyak, 2005), we may reduce the number of hotspots to a more compact size. On the other hand, to do so we may end up with an interval containing many genes and this may bring difficulties in interpretation.

In a similar analysis of the same yeast dataset by Storey *et al.* (2005), the authors also found out the pattern that gene expressions associated with a common SNP marker are tend to be in the same pathway given that the pathway information is available. In their analysis, for instance, twelve expressions were identified to have strong linkage with the SNP marker at one locus on chromosome 3. Two out of these 12 traits are included in the same KEGG pathway: MAPK signaling pathway (pathway id "04010"). Seven expression traits were shown to have linkage with the SNP marker at another locus on chromosome 3. Three out of these 7 traits are in the same pathway: Valine, leucine and isoleucine biosynthesis pathway (pathway id "00290"). These two loci were also detected to be pathway regulators in the current study. These results underscore the importance in finding genetic regulators responsible for the joint expression change of a pathway.

From the biological perspective, due to limited knowledge in genome annotation and gene pathways, not all genes can be mapped to a pathway. In the current analysis, only 1193 gene expressions are mapped to the 99 KEGG pathways, leaving a large proportion of genes unmapped. As an alternative, one can focus the analysis on Gene Ontology (GO) terms which have a more comprehensive coverage of the gene information. As more and more gene information is documented in the public database (e.g., KEGG), this will eventually not being an issue. With limited pathway information, we can also classify gene expressions according to their correlation information to construct gene co-expression networks or modules (Dong and Horvath, 2007). These modules can be treated as pseudo pathways for further analysis. When a module is found to be significantly regulated, the function of those unknown genes can thus be inferred from the genes with known function in the same module. Since genes in the same module or network potentially share the same regulator, such studies can generate meaningful biological hypothesis for experiment test. In addition, principle component analysis can also be applied to reduce the dimension of the original data. Analysis can then be focused on the reduced dimension. It should be noted that the proposed method cannot be applied to substitute the univariate analysis. It rather should be applied to complement the single trait analysis to identify factors with small marginal but strong joint effects.

Acknowledgement

Reference

Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L.V., de Haan, G., Su, A.I., Jansen, R.C. (2008). Genetical genomics: spotlight on QTL hotspots. *PloS Genet. 4*(10): e1000232.

Brem, R.B., Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci.* 102(5): 1572-1577.

Brem, R.B., Yvert, G., Clinton, R., Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296: 752-755.

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K.Y., Morley, M., Spielman, R.S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33: 422–425.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10: 184-194.

Dong, J., Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst. Biol.* 1: 24.

Frary, A., Nesbitt, T.C., Grandillo, S., Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K.B., Tanksley, S.D. (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289: 85-88.

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

Li, C.B., Zhou, A.L., Sang, T. (2006). Rice domestication by reducing shattering. *Science* 311: 1936-1939.

Johnson, R.A., Wichern, D.W. (2007). *Applied multivariate statistical analysis*, New Jersey: Prentice Hall.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., uigserver, P., Carlsson, E., Ridderstrale, 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**:** 267-273.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743-747.

Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L. (2007). Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat. Genet.* 39(4): 496-502.

Schadt, E.E., Lamb, J., Yang, X., *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37:710-717.

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

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

Sun, W. (2007). Statistical Strategies in eQTL Studies. PhD thesis. University of California at Los Angeles.

Wang, K., Li, M.Y., Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81: 1278-1283.

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

Wu, C. L., 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., Su, A.I. (2008). Gene set enrichment in eQTL data identifiers novel annotations and pathway regulators. *PLoS Genet.* 4: e1000070.

Zapala, M.A., Shork, N. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci.* 103: 19430-19435.

Supplemental information — Tables

Suppl. Table 1: List of 99 KEGG pathways and their ID numbers.

#	PID	Funcion	#	PID	Funcion
1	04010	MAPK signaling pathway	2	00460	Cyanoamino acid metabolism
3	00780	Biotin metabolism	4	00910	Nitrogen metabolism
5	00280	Valine, leucine and isoleucine degradation	6	00410	Beta-Alanine metabolism
7	00730	Thiamine metabolism	8	00230	Purine metabolism
9	00550	Peptidoglycan biosynthesis	10	00500	Starch and sucrose metabolism
11	00190	Oxidative phosphorylation	12	00640	Propanoate metabolism
13	03020	RNA polymerase	14	00960	Alkaloid biosynthesis II
15	00051	Fructose and mannose metabolism	16	00052	Galactose metabolism
17	03022	Basal transcription factors	18	00053	Ascorbate and aldarate metabolism
19	04070	Phosphatidylinositol signaling system	20	00290	Valine, leucine and isoleucine biosynthesis
21	00740	Riboflavin metabolism	22	00240	Pyrimidine metabolism
23	00380	Tryptophan metabolism	24	00510	N-Glycan biosynthesis
25	00010	Glycolysis / Gluconeogenesis	26	00330	Arginine and proline metabolism
27	00650	Butanoate metabolism	28	03030	DNA replication
29	00970	Aminoacyl-tRNA biosynthesis	30	00600	Sphingolipid metabolism
31	00790	Folate biosynthesis	32	00920	Sulfur metabolism
33	00100	Biosynthesis of steroids	34	04111	Cell cycle – yeast
35	00561	Glycerolipid metabolism	36	00061	Fatty acid biosynthesis
37	00562	Inositol phosphate metabolism	38	00563	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
39	00513	High-mannose type N-glycan biosynthesis			
41	00565	Ether lipid metabolism	40	00564	Glycerophospholipid metabolism
43	00520	Nucleotide sugars metabolism	42	04120	Ubiquitin mediated proteolysis
45	00340	Histidine metabolism	44	00020	Citrate cycle (TCA cycle)
47	00480	Glutathione metabolism	46	00980	Metabolism of xenobiotics by cytochrome P450
49	00750	Vitamin B6 metabolism	48	00430	Taurine and hypotaurine metabolism
51	00521	Streptomycin biosynthesis	50	00071	Fatty acid metabolism
53	00072	Synthesis and degradation of ketone bodies	52	00251	Glutamate metabolism
55	04130	SNARE interactions in vesicular transport	54	00252	Alanine and aspartate metabolism
57	00350	Tyrosine metabolism	56	00030	Pentose phosphate pathway
59	03050	Proteasome	58	00670	One carbon pool by folate
61	00300	Lysine biosynthesis	60	02010	ABC transporters - General
63	00120	Bile acid biosynthesis	62	00620	Pyruvate metabolism
65	00760	Nicotinate and nicotinamide metabolism	64	00440	Aminophosphonate metabolism
67	00710	Carbon fixation	66	00260	Glycine, serine and threonine metabolism
69	00624	1- and 2-Methylnaphthalene degradation	68	00530	Aminosugars metabolism
71	00627	1,4-Dichlorobenzene degradation	70	00625	Tetrachloroethene degradation
73	00310	Lysine degradation	72	04140	Regulation of autophagy
75	00630	Glyoxylate and dicarboxylate metabolism	74	03010	Ribosome
77	00450	Selenoamino acid metabolism	76	00130	Ubiquinone biosynthesis
79	00900	Terpenoid biosynthesis	78	00770	Pantothenate and CoA biosynthesis
81	00590	Arachidonic acid metabolism	80	00400	Phenylalanine, tyrosine and tryptophan biosynthesis
83	00220	Urea cycle and metabolism of amino groups			
85	00040	Pentose and glucuronate interconversions	82	00720	Reductive carboxylate cycle (CO2 fixation)
87	01030	Glycan structures - biosynthesis 1	84	00860	Porphyrin and chlorophyll metabolism
89	03060	Protein export	86	00360	Phenylalanine metabolism
91	00271	Methionine metabolism	88	00680	Methane metabolism
93	00361	Gamma-Hexachlorocyclohexane degradation	90	02021	Two-component system - Organism-specific
95	00632	Benzoate degradation via CoA ligation	92	00401	Novobiocin biosynthesis
97	00362	Benzoate degradation via hydroxylation	94	01031	Glycan structures - biosynthesis 2
99	00903	Limonene and pinene degradation	96	00272	Cysteine metabolism
			98	01032	Glycan structures - degradation

Suppl. - Table 2: Detailed list of hotspot regulations

Ch	Hotspot (Gene)	Start	Stop	Expression pathways being regulated
2	YBR131W(CCZ1)	499012	499012	"03020" "00010" "00600" "00790" "00100" "00020" "00300" "00620" "00310" "00900" "00720" "00220"
2	YBR132C(AGP2)	499889	499895	"03020" "00740" "00010" "00600" "00790" "00100" "00300" "00620" "00310" "00900" "00400" "00590" "00220"
2	gBR07	506661	508843	"00500" "03020" "00740" "00010" "00790" "00100" "04111" "00020" "00300" "00620" "00900" "00400" "00590" "03060"
2	YBR139W	516889	517123	"00500" "04070" "00740" "00790" "00100" "04111" "00564" "00620" "00760" "00900" "00400" "00220"
2	gBR08	519049	521415	"00500" "03020" "00740" "00790" "00100" "00564" "00020" "00620" "00900" "00400" "00220" "03060"
2	YBR142W(MAK5)	530481	530481	"00500" "03020" "00740" "00970" "00600" "00790" "00100" "00020" "00620" "00450" "00900" "00400" "00590" "00220" "03060"
2	YBR147W	537314	537314	"00500" "03020" "00740" "00510" "00970" "00600" "00790" "00100" "04111" "00564" "00020" "00620" "00310" "00450" "00900" "00400" "00220" "03060" "00361"
2	YBR154C(RPB5)	548401	548401	"04010" "00500" "03020" "04070" "00510" "00970" "00600" "00790" "00100" "04111" "00562" "00513" "00564" "00020" "00620" "00760" "00450" "00900" "00400" "00590" "00220" "01030" "03060"
2	YBR156C(SLI115) NBR031C NBR034W NBR035W	551299 553812 555575 555778	551299 553812 555596 555787	"00500" "00640" "03020" "04070" "00740" "00510" "00970" "00600" "00790" "00100" "04111" "00562" "00513" "00564" "00020" "00030" "00620" "00760" "00450" "00900" "00400" "00590" "00220" "01030" "03060
2	YBR161W	562409	562415	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "04111" "00513" "00564" "00020" "00620" "00530" "00450" "00900" "00400" "00220" "00360" "03060"
2	YBR162W(YSY6)	565216	565216	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "04111" "00564" "00030" "00620" "00530" "00450" "00900" "00400" "00220" "00360"
2	YBR163W(DEM1) YBR165W(UBS1)	567221 569414	567221 569414	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "00564" "00030" "00620" "00900" "00400" "00220" "01030"
2	YBR165W(UBS1) YBR166C(TYR1)	569420 570229	569420 570229	"00500" "03020" "00650" "00600" "00790" "00100" "00564" "00030" "00620" "00900" "00400" "00220" "01030"
2	YBR172C(SMY2) YBR174C	579459 582419	579459 582419	"00500" "00960" "04070" "00650" "00600" "00790" "00100" "04111" "00562" "00564" "00900" "00400" "00220"
2	YBR176W(ECM31) NBR038W NBR041W	584351 592863 592989	584357 592863 592989	"00500" "00960" "00650" "00600" "00790" "0100" "00564" "00620" "00530" "00900" "00400" "00220"
3	YCL026C(FRM2)	75021	75021	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00330" "00650" "00750" "00670" "00630" "00770" "00680" "00401" "00903"
3	YCL025C(AGP1)	76127	76127	"00910" "00280" "00410" "00640" "00053" "00290" "00740" "00380" "00010" "00330" "00650" "00561" "00020" "00340" "00750" "00071" "00252" "00670" "00300" "00620" "00120" "00310" "00630" "00770" "00220" "00360" "00680" "00401" "00272" "00903"
3	YCL023C	79091	79091	"04010" "00910" "00280" "00410" "00640" "00053" "00290" "00740" "00380" "00010" "00330" "00650" "00600" "00920" "04111" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00624" "00310" "00630" "00450" "00770" "00720" "00220" "00360" "00680" "00401" "00272" "00903"
3	YCL022C gCL01 YCL018W(LEU2)	81832 90412 91977	81832 91496 92391	"04010" "00910" "00280" "00410" "00640" "00051" "00053" "00290" "00380" "00010" "00330" "00650" "00970" "00600" "00920" "04111" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00440" "00260" "00624" "00310" "00630" "00450" "00770" "00720" "00220"

				"00360" "00680" "00271" "00401" "00632" "00272" "00903"
3	YCL014W(BUD3)	100213	100213	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00330" "00650" "00600" "00920" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00260" "00310" "00630" "00450" "00770" "00220" "00680" "00401" "00272" "00903"
3	YCL009C(ILV6)	105042	105042	"04010" "00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00330" "00650" "00920" "00561" "00563" "00564" "00020" "00340" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00260" "00710" "00310" "00630" "00450" "00770" "00220" "00360" "00680" "00401" "00272" "00903" "00600"
3	NCR015C gCR02	175799 177850	175808 177850	"04010" "00280" "00410" "00290" "00380" "00600" "00563" "00340" "00350" "00440" "00630" "00450" "00770" "00903"
5	gEL02 YEL021W(URA3) NEL011C YLR014C(PPR1)	109310 116530 117046 117705	109310 116830 117056 117705	"00740" "00240" "00510" "00020" "00251" "00252" "00030" "00620" "00710" "00630" "00720" "01030"
12	YLR236C	611967	611997	"00410" "00380" "00650" "00600" "00100" "00513" "00071" "00072" "00620" "00530" "00627" "00900" "00720" "00361" "00903"
12	gLR07	634225	634226	"00280" "00640" "00380" "00650" "00600" "00100" "00072" "04130" "00620" "00530" "00627" "00900" "00720" "01030" "00361" "00903"
12	gLR07	634227	634227	"00280" "00640" "00380" "00650" "00600" "00100" "00513" "00072" "04130" "00620" "00530" "00627" "00310" "00900" "00720" "01030" "00361" "00903"
12	gLR07	635380	635380	"00280" "00410" "00190" "00640" "00240" "00380" "00010" "00650" "00600" "00100" "00564" "00071" "00072" "04130" "00620" "00530" "00627" "00310" "00900" "00720" "00860" "01030" "00361" "00903"
12	YLR252W YLR253W	642137 644082	642137 644136	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00072" "04130" "00620" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "00361" "00903"
12	YLR257W YLR258W(GSY2)	659357 672779	659357 672785	"00280" "00410" "00190" "00640" "00960" "00240" "00380" "00510" "00010" "00650" "00600" "00100" "00561" "00513" "00564" "00020" "00071" "00072" "00252" "04130" "00030" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "03060" "00361" "00632" "00903"
12	YLR261C(VPS63) YLR263W(RED1)	668249 672779	668249 672785	"00280" "00410" "00190" "00640" "00240" "00380" "00510" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00252" "04130" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "03060" "00271" "00361" "00903"
12	YLR265C(NEJ1)	674651	674651	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00252" "04130" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "03060" "00361" "00903"
12	NLR16W	677957	677957	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00252" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "00361" "00903"
12	YLR267W(BOP2) YLR269C YLR271W	679808 681096 683457	679808 681096 683463	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00252" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "00361" "00903"
12	YLR273C(PIG1) YLR274W(CDC46)	689211 693610	689217 693616	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00620" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "00271" "00361" "00903"
12	YLR274W(CDC46)	693790	693790	"00280" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00620" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "00271" "00361" "00903"
12	YLR277C(YSH1)	697260	697260	"00280" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00620" "00530" "00627" "00130" "00900" "00720" "00860"

				"00271" "00361" "00903"
12	YLR281C	704828	704828	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00564" "00071" "00072"
	YLR282C	705088	705220	"00252" "00620" "00627" "00900" "00720" "00271" "00361"
12	YLR282C	705226	705226	"00280" "00640" "00380" "00650" "00600" "00100" "00072" "00252" "00620" "00627" "00130" "00900" "00720" "00361"
12	YLR285W	708035	708041	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00020" "00071" "00072"
	NLR118C	708258	708260	"00620" "00627" "00130" "00900" "00720" "00361"
	YLR286C(CTS1)	708594	708594	
12	NLR121W	710924	710924	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00020" "00072" "00620" "00627" "00130" "00900" "00720" "00361" "00361"
12	YLR288C(MEC3)	713638	713644	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00564" "00071" "00251" "00072" "00252" "00620" "00627" "00130" "00900" "00720" "00271" "00361"
12	YLR288C(MEC3)	713686	713686	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00071" "00072" "00620" "00627" "00130" "00900" "00720" "00361"
12	YLR292C(SEC72)	719857	719857	"00280" "00640" "00010" "00650" "00600" "00100" "00072" "00627" "00450" "00900" "00720" "00271" "00361"
13	YML120C(NDI1)	28622	28694	"00280" "00410" "00640" "00053" "00290" "00380" "00561" "00071" "00710" "00770" "00220" "00272" "00903"
13	NML013W	46070	46084	"00280" "00410" "00640" "00053" "04070" "00290" "00380" "00330" "00650" "00970" "00561" "00020" "00071" "00251" "00670" "00120" "00260" "00710" "00310" "00770" "00220" "00272" "00903"
13	NML011W	49894	49903	"00910" "00280" "00410" "00500" "00640" "00053" "04070" "00290" "00380" "00010" "00330" "00650" "00970" "00561" "00020" "00071" "00251" "00252" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00720" "00220" "00272" "00903"
13	YML108W	54913	54913	"00280" "00410" "00640" "00053" "00290" "00380" "00650" "00970" "00561" "00020" "00071" "00251" "00670" "00120" "00260" "00710" "00770" "00220" "00272" "00903"
	YML106W(URA5)	57145	57145	
13	gML01	64970	69122	"00280" "00410" "00640" "00053" "00290" "00380" "00650" "00970" "00561" "00071" "00251" "00252" 00670" "00120" "00260" "00310" "00770" "00220" "00040" "00272" "00903"
13	YML098W(TAF13)	77684	77684	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00251" "00670" "00120" "00310" "00770" "00220" "00272" "00903"
13	YML097C(VPS9)	78655	78655	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00120" "00770" "00220" "00272" "00903"
13	NML009C	79760	79786	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00710" "00310" "00770" "00220" "00272" "00903"
	YML096W	81250	81358	
13	YML091C(RPM2)	87587	87587	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML086C(ALO1)	96015	96015	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML084W	99585	99720	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00310" "00770" "00360" "00272" "00903"
13	YML083C	100048	100048	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML072C	124876	124876	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00650" "00561"
	YML071C(DOR1)	129925	129925	"00340" "00071" "00670" "00620" "00120" "00710" "00220" "00272" "00903"
13	YML061C(PIF1)	149075	149075	"00280" "00410" "00640" "00053" "00650" "00561" "00340" "00071" "00670" "00120" "00710" "00220" "00272" "00903"
14	NNL035W	449639	449639	"04010" "00410" "00500" "00190" "03020" "00960" "00051" "00052" "03022" "00053" "04070" "00290" "00240" "00380" "00510" "00010" "00330" "00650" "03030" "00970" "00600" "00790" "00061" "00562" "00563" "00564" "00565" "04120" "00520" "00340"

|				
				"00480" "00750" "00071" "00521" "00251" "00252" "04130" "00030" "00350" "00670" "03050" "00300" "00620" "00120" "00440" "00760" "00260" "00710" 00625" "00310" "03010" "00450" "00400" "00590" "00220" "00860" "00040" "00360" "01030" "03060" "00271" "01031" "00632" "00272"
14	YNL074C(MLF3)	486861	486861	"04010" "00500" "00190" "00640" "03020" "03022" "04070" "00290" "00240" "00380" "00510" "00010" "00330" "00650" "03030" "00970" "00790" "00920" "00061" "00562" "00563" "00513" "04120" "00020" "00480" "00750" "00521" "00251" "00252" "04130" "00030" "00350" "00670" "03050" "00300" "00620" "00440" "00760" "00260" "00710" "04140" "00310" "00450" "00400" "00590" "00220" "00860" "01030" "00680" "03060" "00271" "01031" "00632" "00272"
14	YNL066W(SUN4)	502316	502316	"03022" "04070" "00290" "00330" "00970" "00790" "00920" "00562" "00563" "00750" "00251" "00252" "00300" "00620" "00440" "00760" "00260" "00310" "00450" "00400" "00590" "00220" "00860" "00271" "00632" "00272"
14	gNL07	525061	525064	"03022" "04070" "00290" "00330" "00970" "00562" "00563" "00750" "00251" "00252" "00300" "00620" "00760" "00260" "00630" "00400" "00590" "00220" "00860" "00271"
15	YOL094C(RFC4)	141621	141633	"04010" "00052" "04111" "00561" "00513" "00564" "00565" "00980" "00071" "00030" "00350" "00120" "00710" "00624" "04140" "00130" "00040"
15	YOL093W YOL092W YOL089C(HAL9)	143597 144659 150651	143597 144659 150651	"00561" "00513" "00565" "00980" "00071" "00251" "00252" "00030" "00350" "00120" "00710" "00624" "00625" "04140" "00040"
15	YOL088C(MPD2)	154177	154309	"04010" "00561" "00513" "00564" "00565" "00520" "00980" "00071" "00251" "00350" "00120" "00710" "00624" "00625" "04140" "00040"
15	gOL02	170945	174364	"04010" "00910" "00280" "00410" "00500" "00640" "00053" "00380" "00010" "00790" "00920" "00561" "00513" "00564" "00565" "00520" "00340" "00980" "00430" "00750" "00071" "00521" "00251" "00252" "00030" "00350" "00670" "00120" "00760" "00260" "00710" "00624" "00625" "00590" "00860" "00040" "00360" "00680" "00272" "00362" "00903"
15	gOL02 YOL081W(IRA2) YOL080C(REX4)	175594 180180 180961	179289 180222 180961	"04010" "00910" "00280" "00410" "00640" "00053" "00380" "00010" "00790" "00920" "00561" "00513" "00564" "00565" "00520" "00340" "00980" "00430" "00750" "00071" "00521" "00251" "00252" "00030" "00350" "00670" "00120" "00760" "00260" "00710" "00530" "00624" "00625" "04140" "00310" "00590" "00220" "00860" "00040" "00360" "00680" "00271" "00272" "00362" "00903"
15	YOR127W(RGA1)	563943	563943	"00052" "00380" "00010" "00561" "00020" "00980" "00071" "00350" "03050" "00120" "00710" "00624" "00130"

Supplemental information–Figures
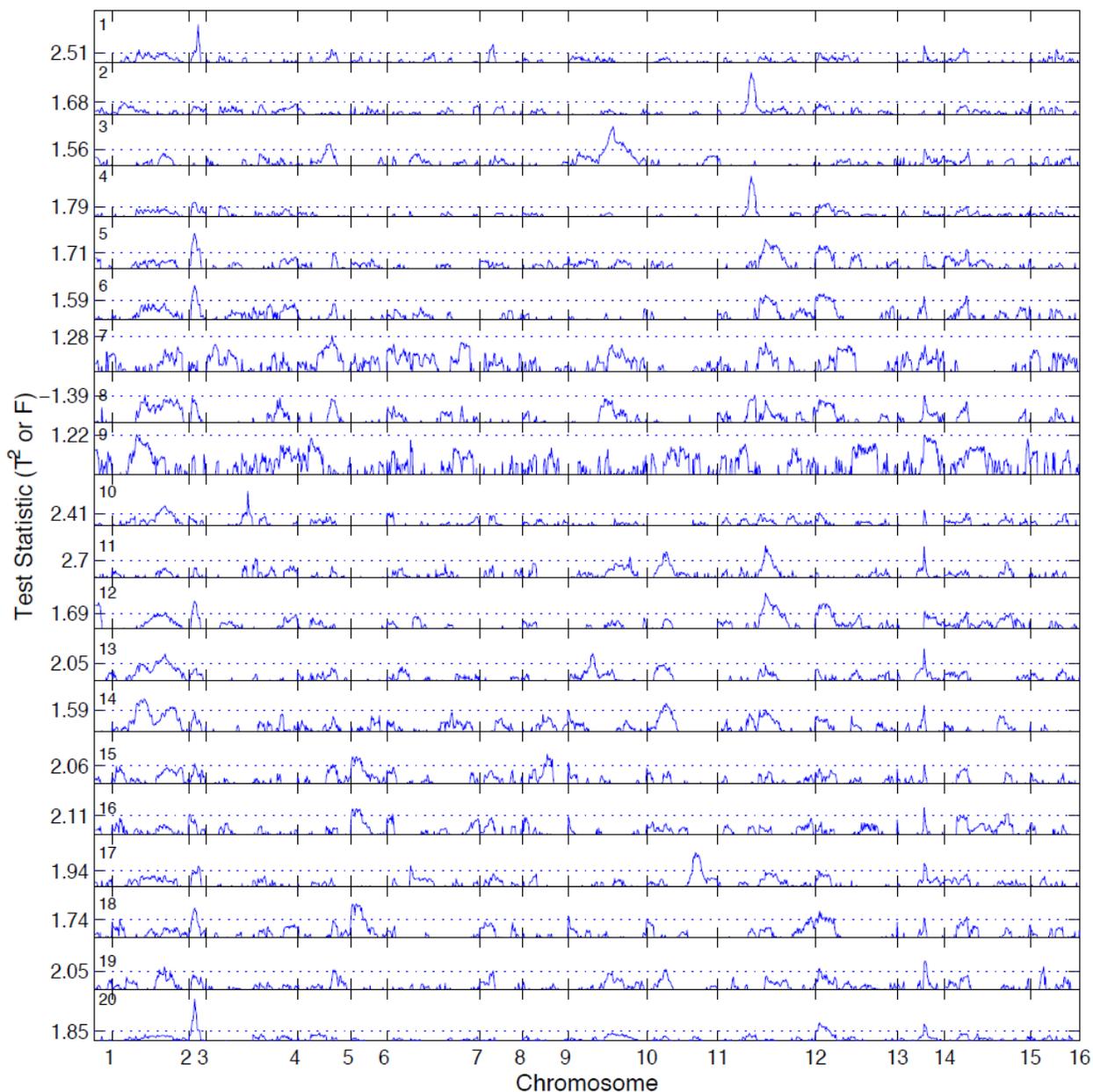
Suppl. - Figure 1: The $\log_{10}$ based T^2 (or F) statistic values across the entire 16 chromosomes of the yeast genome for EPs 1-20. Dotted horizontal line indicates the 5% genome-wide permutation cutoff. Genomic positions where T^2 (or F) values pass the threshold harbor potential pathway regulators for the corresponding EP. See Suppl. - Table 1 for the pathway information.
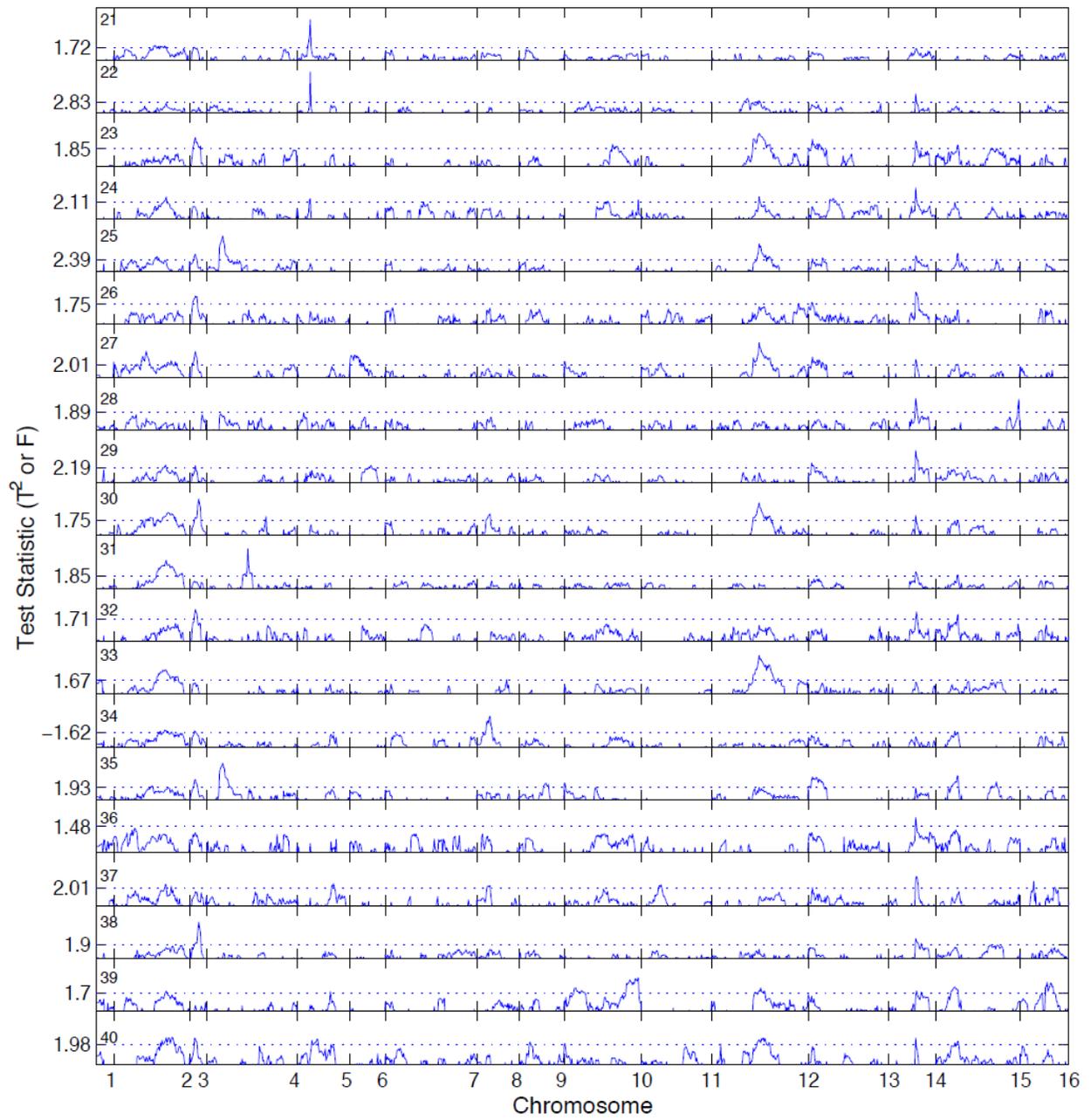
Suppl. - Figure 1 continued.

Suppl. - Figure 1 continued.

23

Suppl. - Figure 1 continued.

24
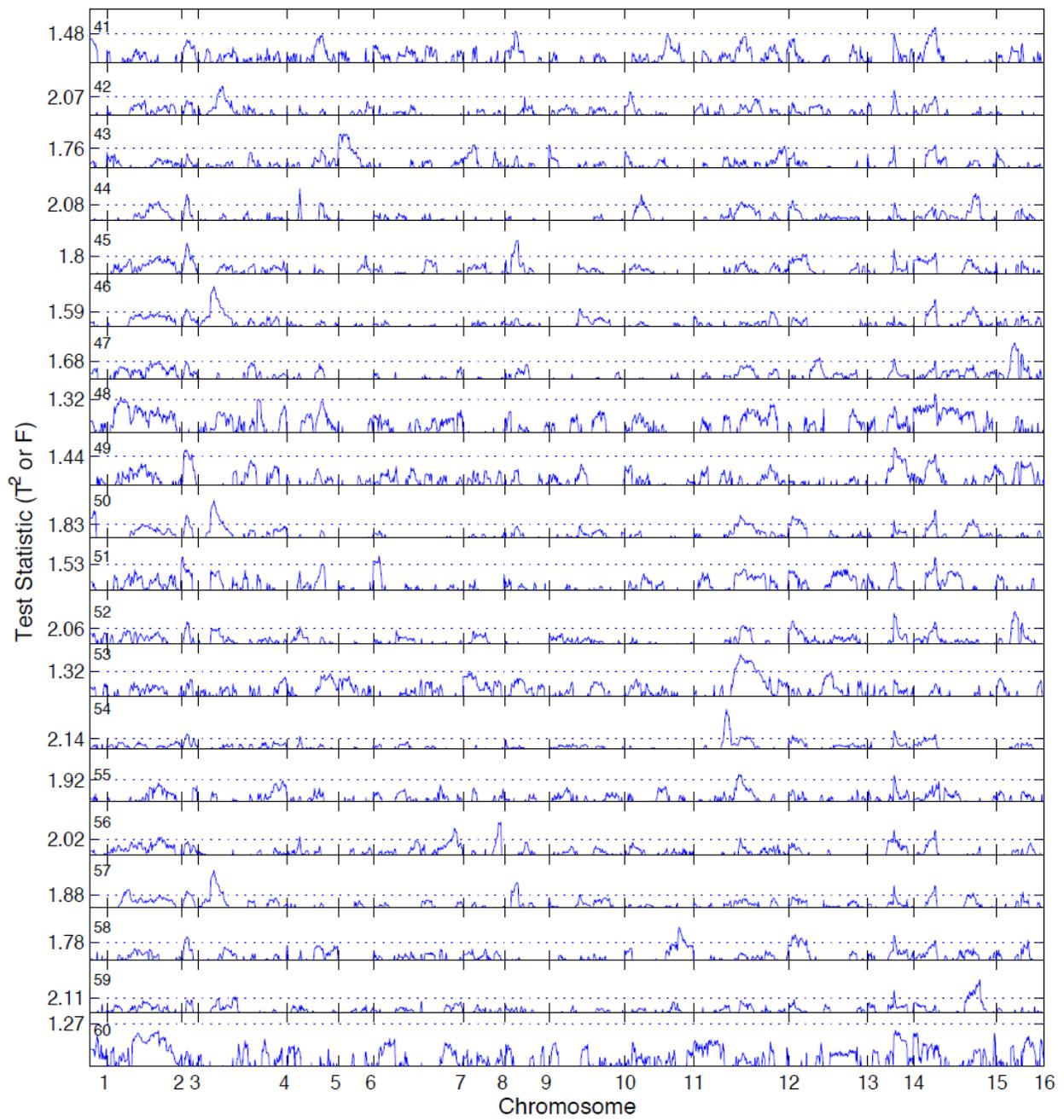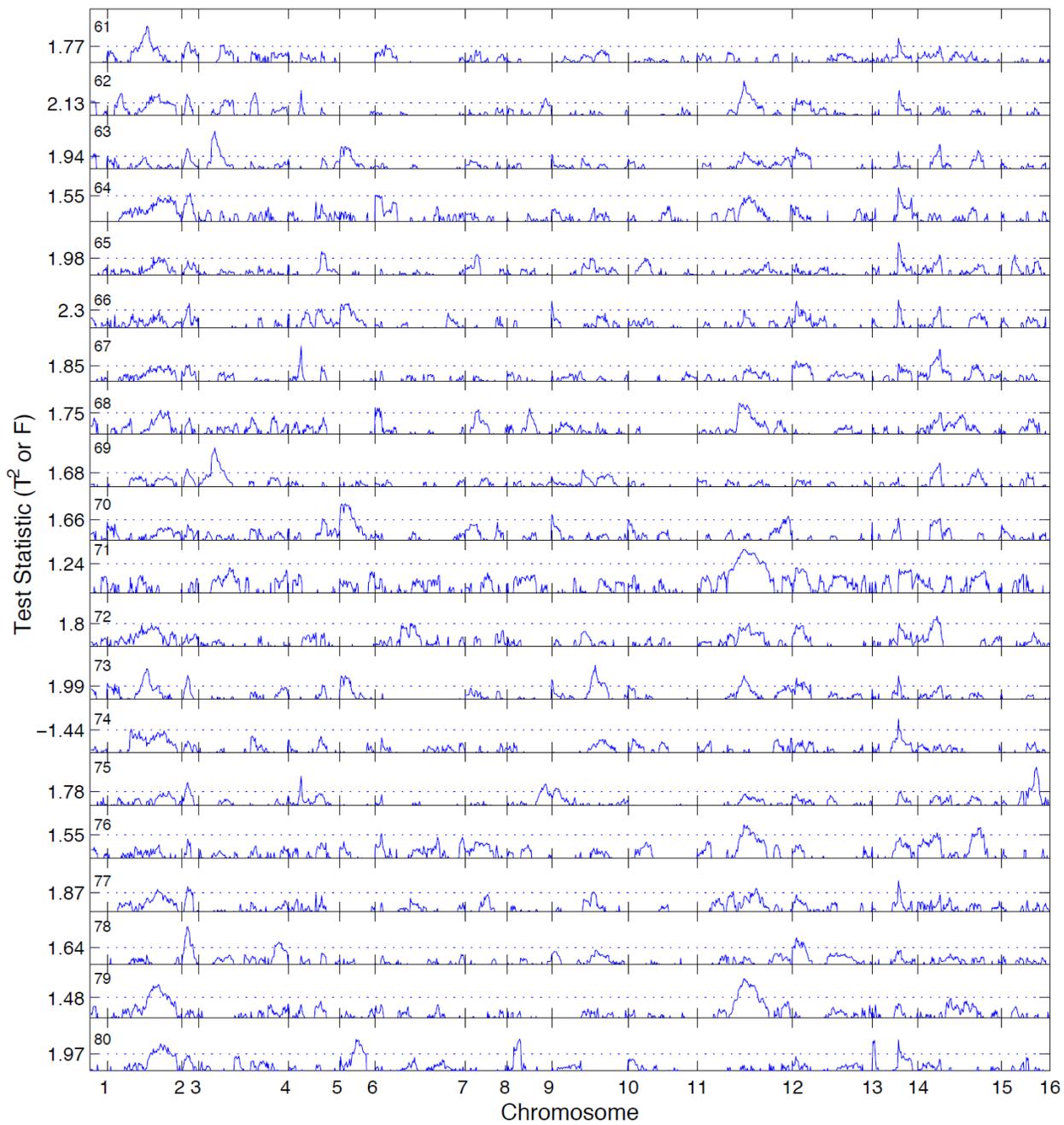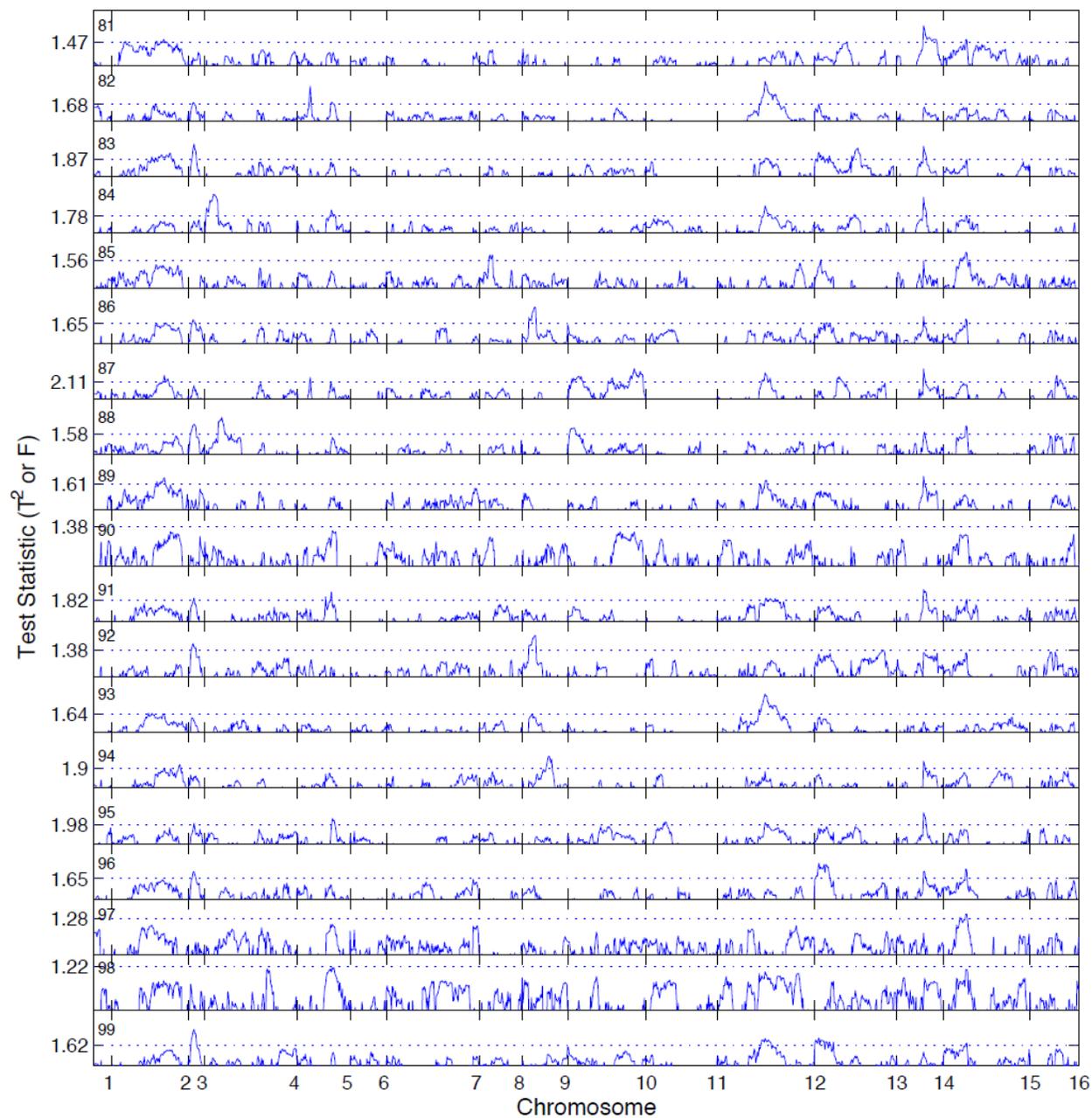
Suppl. - Figure 1 continued.
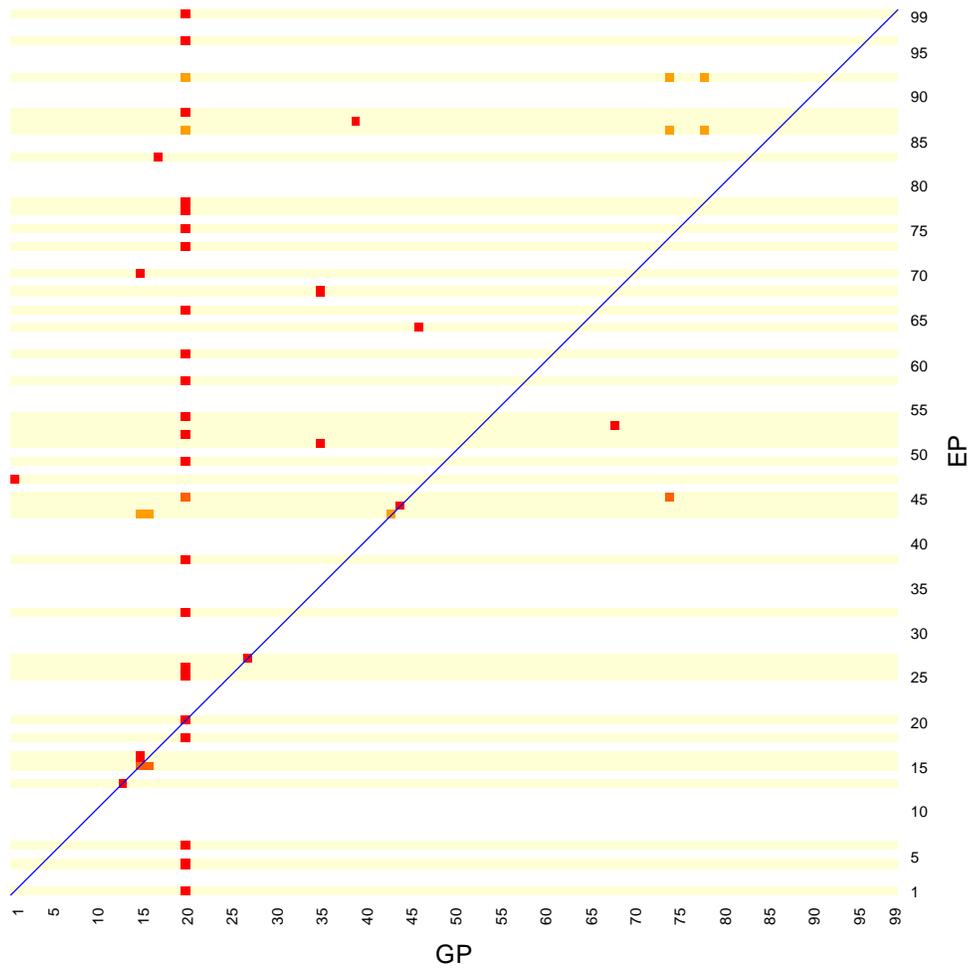
Suppl. - Figure 2: A heatmap of enriched pathways. Only significantly enriched pathways are shown in the plot (indicated by squares). Squares on the diagonal line indicate *cis*-pathway regulation and those on off-diagonals indicate *trans*-regulation. The horizontal and vertical axes denote the genetic pathway (GP) and the gene expression pathway (EP), respectively. Strong *trans*-pathway regulations are detected.