Gene Expression Level Association Analysis for Yeast Data

Group 5

Liu Huihang	SA18017026
Wu Jing	SA18017054
Chen Ziyu	SA18017046
Xu Tianchen	SA18017038

1. Introduction

Gene expression levels are fundamental for differences of phenotypic traits among a particular species. SNPs that regulate mRNA expression of a gene are so-called expression Quantitative Trait Loci (eQTL) (Schadt et al., 2003). Gene expression levels are treated as quantitative traits for the identification of eQTLs that lead to differences in expression levels. The study of the relationship between eQTLs and gene expression may be complex due to the presence of both local and distant genetic effects and shared genetic components across multiple genes Brem and Kruglyak (2005); Cai et al. (2012). So our biological problem is to study the association between eQTLs and gene expression levels. Yeast is one of the simplest eukarya with a short life cycle and small genome size, of which growth can be controlled easily by the experimenter. The yeast data is completely sequenced data so that we have enough information to find the association between SNPs and gene expression. Besides, the previous study has discovered that yeast cells share many similar genes with human cells, which means examining genes of yeast helps to learn the roles of them in human disease. Therefore yeast can be a model organism for gene expression study.

In a yeast expression study, researchers often conduct the eQTLs analysis to understand how the SNPs influence the expression level of genes in the yeast MAPK signaling pathways. Some genetic and biochemical analysis has revealed that there are a few functionally distinct signaling pathways of genes (Gustin et al., 1998; Brem and Kruglyak, 2005), suggesting that the association structure between the eQTLs and the gene is of low rank, and the pattern of sparsity should be pathway-specific. Furthermore, each signaling pathway involves only a subset of genes, which are regulated by only a few genetic variants, suggesting that each association between the eQTLs and the genes is sparse in both the input and the output.

From the statistical perspective, we treat gene expression levels as responses and genotypes at SNPs as predictors. Then we convert our biological problem to a multivariate regression problem. And we except sparsity in the regression coefficient matrix in the biological context. Therefore, variable selection is required here, which is precisely what we do in our project.

The rest of the paper is organized as follows. Section 2 introduces the Yeast data we used and illustrates how we process them. Section 3 introduces three different methods we used and discusses their performence. We summarize our results in Section 4 by combining the results from both statistical and biological aspects. An associated R package implement-

ing the suggested method is available at `https://github.com/huihangliu/GWAS/tree/master/Final`.

2. Yeast Data

2.1 Data Describtion

We explored the effects of different statistical methods by by the analysis of a yeast eQTL data set described by Brem and Kruglyak (2005), where $n = 112$ segregants were grown from a cross between two budding yeast strains, BY4716 and RM11-1a. For each of the segregants, gene expression was profiled on microarrays containing 6216 genes, and genotyping was performed at 2957 markers.

Genotype is a categorical variables with a value of 0 or 1, and gene expression level is given by $log2$(sample/BY reference).

2.2 Data Preparation

2.2.1 Processing Markers Data

The DNA of the same organism has a very high similarity. For example, all humans are 99.9% identical and, of that tiny 0.1% difference. Considering the similarity of markers, we found that different markers in the data may have identical or differ at most few yeast samples after exploratory data analysis.

Inspired by Yin and Li (2011), we combined the markers into 949 blocks such that markers with the same block differed by at most one sample, and one representative marker was chosen from each block. To accomplish that, we used hierarchical clustering with complete-linkage criteria and cut the clustering tree at distance $d = 1$. Note that the value of the marker data is only 0 or 1, so we can use any distance function, like Euclidean distance or Manhattan distance, to obtain the same result.

We can further reduce the number of markers, observing that different markers have different functions, and some markers have no effect on the gene expression we want to study. Thus a marginal gene–marker association analysis was then performed to identify markers that are associated with the expression levels of at least two genes with a p-value less than 0.05, resulting in a total of $p = 776$ markers with additional 18.2% reduction.

2.2.2 Processing Expression Level Data

We choose genes according to MAPK signaling pathways (Kanehisa et al., 2013) from *The yeast MAPK pathway from the KEGG database* [1] as shown in Figure 1.

From the database we obtain 53 markers which have been linked to MAPK signaling pathways biologically.

After the above operations, we got the processed data with a $X \in \mathbb{R}^{112\times949}$, $Y \in \mathbb{R}^{112\times53}$.

1. http://www.genome.jp/kegg/pathway/sce/sce04011.html

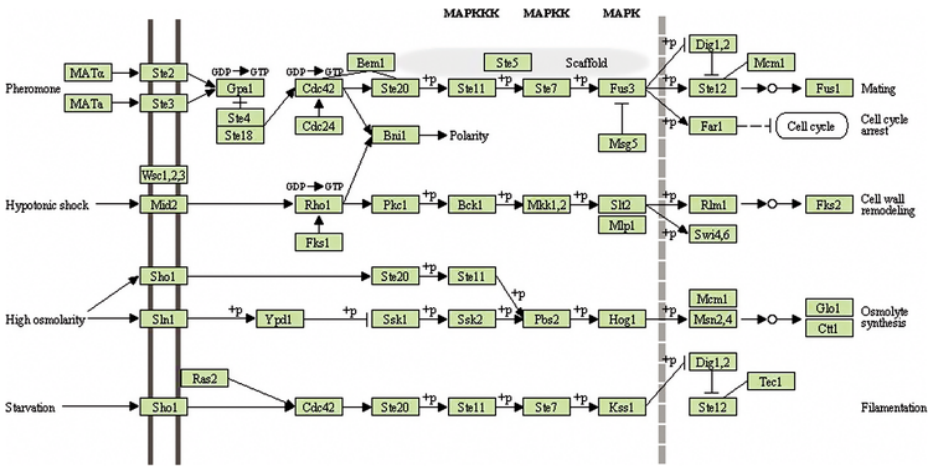

Figure 1: The yeast MAPK pathway from the KEGG database

Hereinafter, we denote p as the number of explanatory variables, q as the number of response variable, n as the number of samples, E as the random error matrix, and B as the coefficient matrix, we can construct a multi-response linear model as following

$$Y = XB + E. \tag{1}$$

3. Methodology

Here we are going to introduce methods we used to select variables.

As described in (1), we are facing a linear regression model in high dimensional space with multiple response variables. This makes it more difficult for us to deal with problems.

Next, we will introduce three methods to solve the model. They were verified as a good method to deal with univariate high-dimensional linear regression, grouped variables, and factorization problem.

3.1 Unit-Response Regression

Denote Y_j as the j-th colume of Y, which represents the expression level of j-th gene.

It is intuitive for us to divide and conquer it by regressing each Y_j with X and and we will get q linear models.

We can use classical LASSO (Tibshirani, 1996) to get the estimated coefficient vector $\hat{\beta}_{\cdot j} \in \mathbb{R}^p$. Then combine q coefficient vectors $\hat{\beta}_{\cdot j}$ into a matrix $\widehat{B} \in \mathbb{R}^{p\times q}$ by column.

We plot heatmap of predicted Y and the real one in Figure 2 because it is simpler and more intuitive than displaying numbers. We can learn from the diagram that the two colors are almost the same, indicating that the estimation results by LASSO are acceptable.

However, as is known, LASSO is a kind of shinkage estiamtion method with L_1 penalty, which not only shrinks the size of the coefficients, but also set some of them to zero. So a sparse $\hat{\beta}_{\cdot j}$ is expected for each $j \in \{1, 2, \ldots, q\}$. We rewrite the optimazation function as

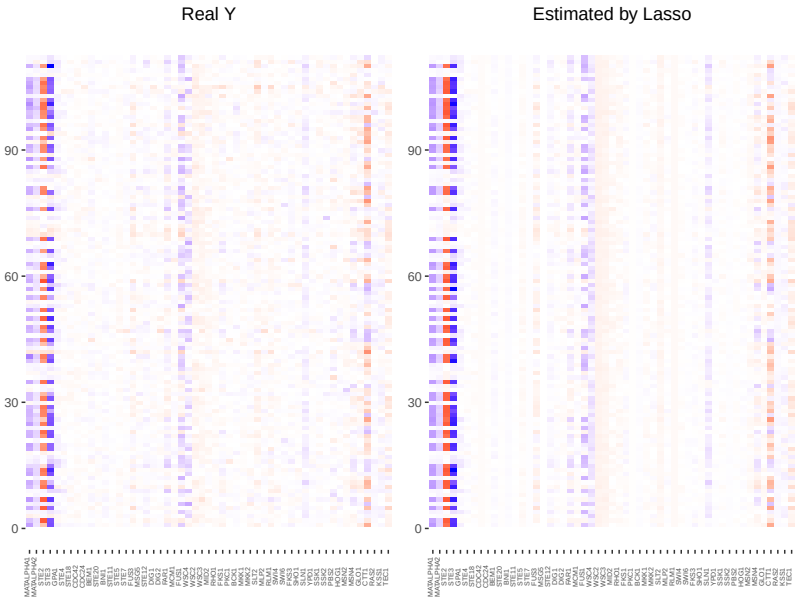

Figure 2: Heatmaps of real Y and $\hat{Y}$ by LASSO

following

$$\widehat{B} = \arg\min_{B} \left\{ \frac{1}{2n} \|Y - XB\|_F^2 + \sum_{i=1}^{q} \lambda_i \|B_{\cdot i}\|_1 \right\} \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\sum_{i=1}^{q} \lambda_i \|B_{\cdot i}\|_1$ can be regraded as a $(1,1)$ norm of matrix B.

Actually, we got 602 nonzero rows in $\widehat{B}$ which also has full column rank.

3.2 Grouped-Response Regression

Group sparse linear regression for multitask learning (Dai and Barber, 2016)

$$\widehat{B} = \arg\min_{B} \left\{ \frac{1}{2n} \|Y - XB\|_F^2 + \lambda \|B\|_{(2,1)} \right\} \tag{3}$$

where the $(2,1)$ norm in the penalty is given by $\|B\|_{(2,1)} = \sum_i \sqrt{\sum_j B_{ij}^2}$.

As shown in the Figure 3, although it totally uses fewer predictors variables than LASSO, but it also restores Y well.

Intuitively, $(2,1)$ norm penalty can be understood as a L_1 penalty imposed on sum of squares of the rows of B. If we choose a large λ, $\widehat{B}$ will contain many zero rows, however the nonzero rows will themselves be dense (not entrywise sparsity). This is because $(2,1)$ norm penalty promotes rowwise sparsity of $\widehat{B}$.

Biologically, the result shows that among all p markers, a total of 201 markers are related to 53 genes. And each marker selected has some effect on each gene in our data.

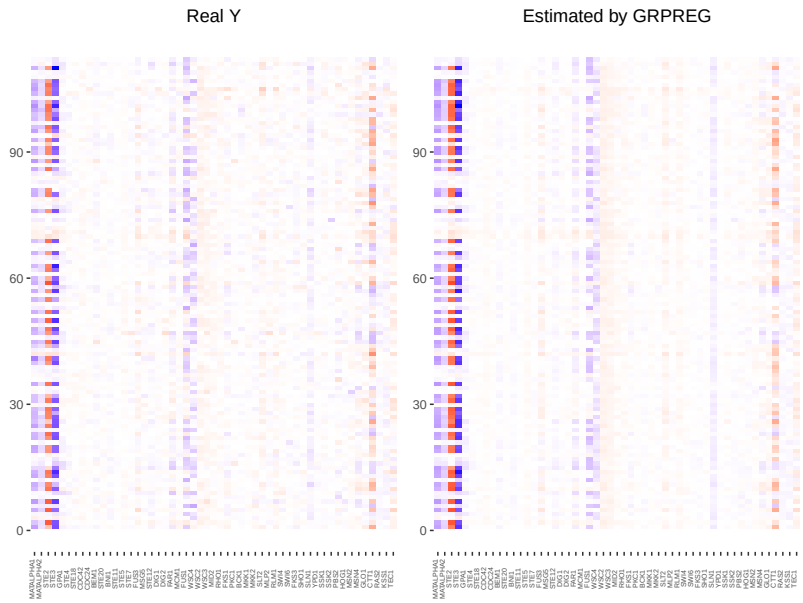

Figure 3: Heatmaps. $\widehat{B}$ has full column rank and 201 non-zero rows.

3.3 Multi-Response Regression

Previous biological research has revealed some facts which could be the guidelines for us to choose suitable method of data analysis.

Extensive genetic and biochemical analysis has revealed that there are a few functionally distinct signaling pathways of genes (Gustin et al., 1998; Brem and Kruglyak, 2005), suggesting that the association structure between the eQTLs and the genes is of low rank. Each signaling pathway involves only a subset of genes, which are regulated by only a few genetic variants, suggesting that each association between the eQTLs and the genes is sparse in both the input and the output (or in both the responses and the predictors), and the pattern of sparsity should be pathway specific. Moreover, it is known that the yeast MAPK pathways regulate and interact with each other (Gustin et al., 1998).

The complex genetic structures described above clearly indecate that the association structure between the eQTLs and the gene is of low rank and sparsity. It call for a joint statistical analysis that can reveal multiple distinct associations between subsets of genes and subsets of genetic variants.

SOFAR (Uematsu et al., 2019) uses the SVD decomposition $B = UDV^T$ and then impose penalties into U, D and V respectively.

$$\begin{aligned}\left(\widehat{D},\widehat{U},\widehat{V}\right) =& \underset{D,U,V}{\arg\min} \left\{ \frac{1}{2n} \left\| X - UDV^T \right\|_F^2 + \lambda_d \|D\|_1 + \lambda_a \rho_a(UD) + \lambda_b \rho_b(VD) \right\} \\ & \text{subject to } U^TU = \mathbf{I}_m, \quad V^TV = \mathbf{I}_m \end{aligned} \tag{4}$$

Rank reduction is achieved mainly through the first term and variable selection is achieved through the last two terms. $\rho_a(\cdot)$ and $\rho_b(\cdot)$ can be equal or distinct, depend-

ing on our questions and goals. ρ can be entrywise L_1 norm or rowwise $(2,1)$ norm or others.

Figure 4 shows the results from SOFAR with turning parameters λ_d, λ_a and λ_b choosen by cross validation. SOFAR returned rowwise sparse singular matrix U and V and a low rank diagram matrix D. It finds only 228 significant markers by selecting non-zero rows in the estimation of U, 25 genes by selecting non-zero rows in the estimation of V and 3 latent patterns within the data.

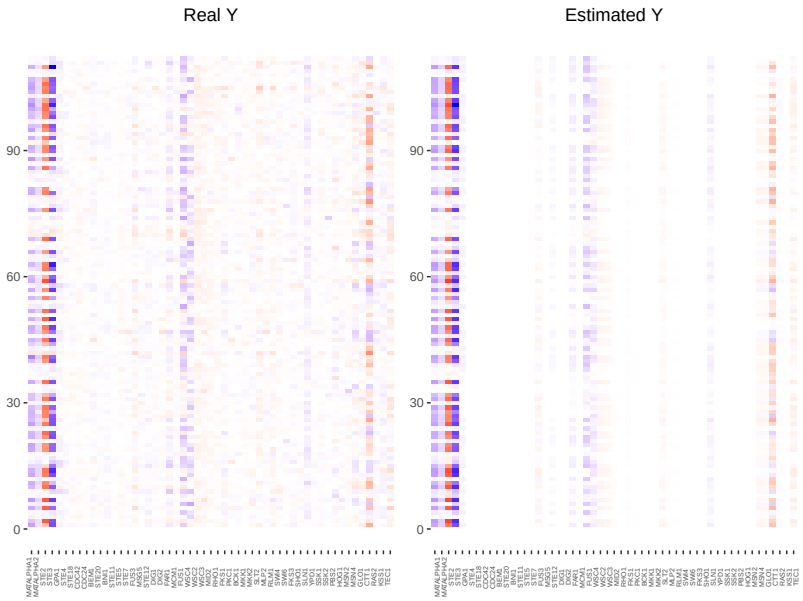

Figure 4: Heatmaps of real Y and $\hat{Y}$ by SOFAR

Specifically, the rank of coefficient matrix is 3, as shown in the Figure 5, indicating that there are only three latent variables (XU). Figure 5 shows the specific linear relationship between XU and response variables YV in different patterns, and the fitting is pretty good. And here we also get the sparsity of U and V.

In order to check whether the SOFAR method completely finds the letent pattern within the data, we drew the fourth possible latent pattern in Figure 6. As is shown, the latent response and the latent predictor are almost completely independent and there is no possible relationship.

4. Summary

Our results clearly demonstrate the sparsity and low-rankness of associations between eQTLs and genes, which allows us to conclude that a certain linear combination of eQTLs has effects only on a subset of genes. Thus, to study a certain group of genes, a reasonable and efficient way is to focus on the corresponding set of eQTLs.

The SVD layers provide more information on samples and genes. The plot for layer 1 indicates that the yeast samples can be divided into two clusters, which means our method can do classification based on the latent variables. As to genes, the nonzero entries in each

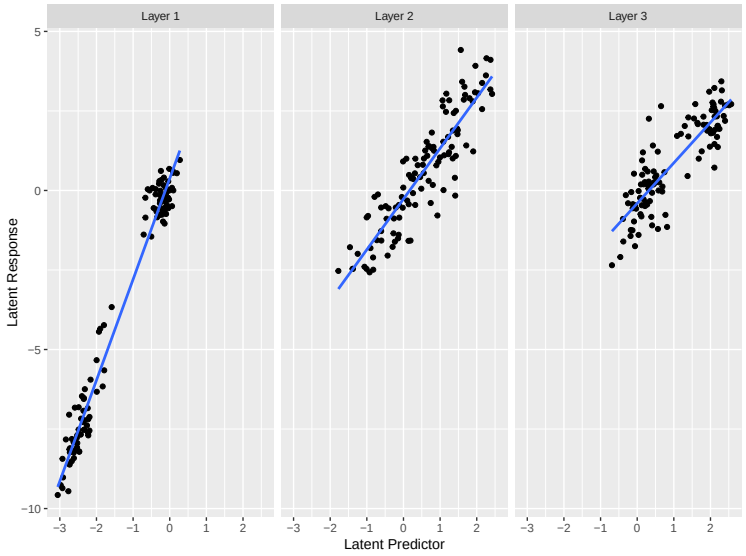

Figure 5: Scatter plots of the latent responses versus the latent predictors in three SVD layers for the yeast data estimated by the SOFAR method

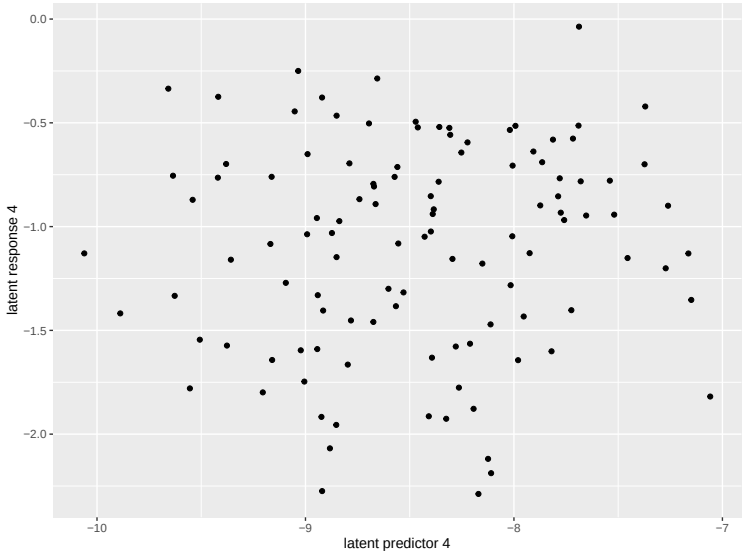

Figure 6: Scatter plots of the 4th latent responses versus the latent predictors for the yeast data estimated by the SOFAR method

column of V match with dominating genes in each layer. The first layer is dominated by four genes, including STE3(-0.69), STE2(0.61), MATα1(-0.32) and MATα2(-0.17). All four genes are upstream in the pheromones pathway, as shown in the top right corner of Figure 1. The second layer include the leading genes CTT1(-0.94), GLO1(-0.14), SLN1(0.14), SLT2(-0.13), MSN4(-0.12) and STE2(-0.11). Notably, MSN4, GLO1, and CTT1 are all downstream genes linked to the downstream gene SLN1 in the high osmolarity pathway.

The leading genes in the third layer can be divided into two groups. The upstream group has STE2(0.26), GPA1(0.20), WSC3(0.19) STE3(0.18) and SLN1(0.15). The downstream group includes FUS1(0.81), FAR1(0.29), STE12(0.11) and TEC1(0.11). They are mainly located in the pheromone and the starvation pathway. Furthermore, STE2 is a leading gene in all three layers. Above all, our results suggest a strong linkage between the upstream and downstream genes, as well as three or four functionally distinct patterns in the MAPK pathway, where each pattern is regulated by a subset of genes. This conclusion is consistent with current findings of genetical studies.

References

Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5): 1572–1577, 2005.

T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2012.

Ran Dai and Rina Foygel Barber. The knockoff filter for FDR control in group-sparse and multitask regression. *arXiv preprint arXiv:1602.03589*, 2016.

Michael C Gustin, Jacobus Albertyn, Matthew Alexander, and Kenneth Davenport. MAP Kinase Pathways in the YeastSaccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.*, 62 (4):1264–1300, 1998.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, 42(D1):D199—-D205, 2013.

Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297, 2003.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Yoshimasa Uematsu, Yingying Fan, Kun Chen, Jinchi Lv, and Wei Lin. SOFAR: large-scale association network learning. *IEEE Transactions on Information Theory*, 2019.

Jianxin Yin and Hongzhe Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics*, 5(4):2630, 2011.