Proposal for Final Project

Group 5

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

1. Question of Interest

(1) How the yeast eQTLs(expression quantitative trait loci), which are regions of the genome containing DNA sequence variants, influence the expression level of genes?

(2) What is the influence of eQTLs on the genes involved in the yeast MAPK signaling pathways?

2. Background

In the genetical genomics experiments of cDNA array of Saccharomyces cerevisiae ORFs, researchers are interested in exploring the relation between the gene expression levels and expression quantitative trait loci (eQTLs) that contribute to phenotypic variation in gene expression. Gene expression levels are usually treated as quantitative traits in order to identify eQTLs.

To maintain a reasonable power given limited sample size and multiple testing correction in eQTL studies, the smallest model with only additive genetic effect is often used to map eQTL (Stranger et al., 2007)

$$y = a + bx + \epsilon$$

where y indicates a gene expression trait and x indicates the additive genetic effect, which can be coded by the number of minor alleles, and ϵ is the residual error. We can easily extend OLS to model the relation bewteen a gene expression and two genetic effects.

To get a more precise result, more genetic effects, especially when the number of genetic effects is less then the observations, need to be modeled simultaneously. OLS fails in the high dimensional linear regression situation. Under some sparsity conditions, many shinkage estimation methods were proposed, for example Tibshirani (1996); Zou and Hastie (2005); Fan and Li (2001). The task can be regarded as a multiple linear regression problem, with the gene expression level as responses and the genetic variants as predictors, as following

$$Y = \mathbf{X}\beta + \epsilon \tag{1}$$

where $Y \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ and $\mathbf{supp}(\beta) = \|\beta\|_0 = s < n \ll p$.

However, the complex genetic structures call for a joint statistical analysis that can reveal multiple distinct associations between subsets of genes and subsets of genetic variants.

Thus, if we treat the genetic variants and gene expressions as the predictors and responses, respectively, in a multivariate regression model, the task can then be carried out by seeking a representation of the coefficient matrix and performing predictor and response selection simultaneously.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$. Some recent methods for eQTL data analysis exploit entrywise or rowwise sparsity of the coefficient matrix to identify individual genetic effects or master regulators (Peng et al., 2010). Dai and Barber (2016) using group structure of variables deduces a group sparse linear regression and Uematsu et al. (2019) suggest the method of sparse orthogonal factor regression via the sparse singular value decomposition with orthogonality constrained optimization.

3. Data Description

The data can be accessed in Gene Expression Omnibus(GEO) by accession number GSE1990. The data were derived from a cross between two strains of the budding yeast: BY4716 and RM11-1a (Brem and Kruglyak, 2005).

Gene expression measurements were obtained for 6216 open reading frames in 112 segregants, and genotypes were identified at 3244 markers.

Title	Genetic complexity in yeast transcripts
Organism	Saccharomyces cerevisiae
Experiment type	Expression profiling by array
Data Size	Data set consists of a 3244×112 **genotype matrix** with 3244 genotypes in rows and 112 samples in columns and a 6216×112 **gene expression matrix** with 6216 genes in rows and 112 samples in columns.
Description	cDNA array of Saccharomyces cerevisiae ORFs. Genotype is category variable, and gene expression level is given by log_2(sample/BY reference)

Table 1: Information about Data

4. Statistical Analysis Plan

(1) *Estimation.* Using some statistical method for example Group lasso (Yuan and Lin, 2006), SOFAR (Uematsu et al., 2019) and SEED (Zheng et al., 2019) to solve the multivariate regression problem (2) with the gene expression levels as responses and the genetic variants as predictors, where both responses and predictors are often of high dimensionality.

(2) *Selection.* Variable (Factor) selection can be achieved by using some shinkage estimation method, actually some method described in estimation procedure can be used to select variables. We will implement some of them.

(3) *FDR Control*. We plan to use knockoff Barber and Candès (2015); Dai and Barber (2016); Candes et al. (2018) to control FDR.

(4) *Conclusion*. Finally, we will compare the results from the methods described above and draw some conclusions from the results given by those methods, especially some biologically significant conclusions.

Note that extensive genetic and biochemical analysis has revealed that there are a few functionally distinct signaling pathways of genes (Brem and Kruglyak, 2005; Gustin et al., 1998), suggesting that the association structure between the eQTLs and the genes is of low rank.

Thus, we choose these sparse multivariate regression (selection) method to complete the plan because they are suitable for the data and explainable when we get a result and very novel to reach unusual (extraordinary) conclusions. See Section 2 for detail.

5. Expected results

After the analysis, we expect to:

- Get a representation of the coefficient matrix β and response selection. And may provide new insights into the complex genetics of gene expression variation.
- Detect power for multiple eQTLs that combine to affect a subset of gene expression traits, which may offer information about the functional grouping structure of the genetic variants and gene expressions.
- Get results which may suggest that there are common genetic components shared by the expression traits of the clustered genes and clear reveal strong associations between the upstream and downstream genes on several signaling pathways, which are consistent with the current functional understanding of the MAPK signaling pathways.

6. Plan B

We may fail in implementing Plan A, because the methods mentioned are novel and the implementation of the plan is challenging.

When our Plan A cannot be implemented, we will

- use some traditional high dimensional linear regression method such as LASSO (Tibshirani, 1996), Elastic-Net (Zou and Hastie, 2005) and SCAD (Fan and Li, 2001) to analysis a single gene expression level by eQTLs, namely expression (1);
- implement some low dimensional method or some old fashioned method such as OLS and BP network to simplified yeast data given by *geneNetBP* package Moharil et al. (2016) in **R**. The data set *yeast* is a data frame of 112 observations of 50 variables: genotype data (genotype states at 12 SNP markers) and phenotype data (normalized and discretized expression values of 38 genes). Both genotypes and phenotypes are of class factor.

We will not fail anymore.

References

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

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.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

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

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

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.

Janhavi Moharil, Maintainer Janhavi Moharil, and Suggests RHugin. Package 'geneNetBP'. 2016.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53, 2010.

Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, and Others. Population genomics of human gene expression. *Nature genetics*, 39(10): 1217, 2007.

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.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Zemin Zheng, M Taha Bahadori, Yan Liu, and Jinchi Lv. Scalable Interpretable Multi-Response Regression via SEED. *Journal of Machine Learning Research*, 20(107):1–34, 2019.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.