An Application of Multivariate Procedure for Genotype Phenotype Mapping in Yeast

Tahir Mehmood*[1], **Solve Sæbø**[1] , **Jonas Warringer**[2] and **Lars Snipen**[1]

[1] Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway

[2] Department of Cell and Molecular Biology, University of Gothenburg, Sweden

Motivation

- Tremendous increase in genomic sequence data
- Demands new or improved methods for exploring the feature construction for genotype-phenotype landscape
-The use of multivariate approaches in genome-wide association studies analysis multivariate relation between genotypes and phenotypes (Mehmood et al. 2011)

Data

-36 Saccharomyces cerevisiae strains were obtained from the Saccharomyces Genome Resequencing Project (SGRP) available at Sanger (http://www.sanger.ac.uk/Teams/Team118/sgrp/).

-6850 protein-coding sequences were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org/)

-Phenotype data: Micro-cultivation of yeast populations during exposure to 2 different treatments (Fructose and Mannose each with 2% concentration)

-Sigmoid growth curves were parameterized into the three fundamental reproductive measures Adaptation, Rate and Efficiency.

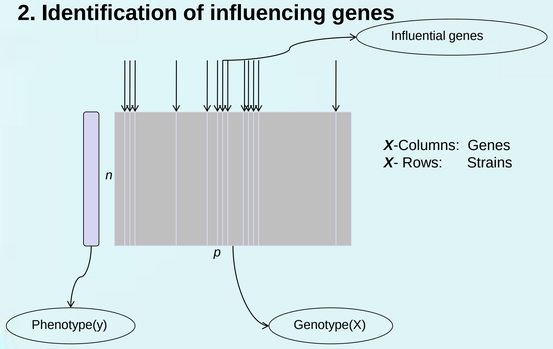
Issues to address

1. Construction of the genotype *X* matrix

For genome *gi* defining the bit-score (tblastx) $S(g_i; r_j)$, indicating to what extent sequence j was found in the respective genome.

Normalizing for length of gene. $N_{i,j} = S(g_i, r_j)/S(r_j, r_j)$

2. Identification of influencing genes



A soft thresholding PLS ST-PLS supervised learning introduced by Sæbø et al. (2007) used for Influential gene selection

Biological relevance

Influencing genes mapped for a phenotype needs to be examined for biological relevance.

Genes can be a priori grouped based on previous studies on S. cerevisiae

- Gene Ontology (GO) variations
- Essential genes
- Genes with known paralogs
- Genes with known frame shift variation
- Genes with known stop codon variation
- Genes with known copy number variations

Enrichment of such categories is an indication of non-random (relevant) detection of influencing genes

Results and Discussion

1. Genotype-phenotype modelling

Labels	Phenotype	No. components	Shrinkage	d-index	No. of genes
Fru_A	Fructose 2%	7	0.7	0.61	50
Fru_R	Fructose 2%	10	0.7	0.75	51
Fru_E	Fructose 2%	10	0.7	0.66	44
Man_A	Mannose 2%	10	0.7	0.59	58
Man_R	Mannose 2%	10	0.7	0.73	43
Man_E	Mannose 2%	9	0.7	0.70	54

Table 1 - Over all distribution of model parameters and performance
Results obtained from the 6 ST-PLS model fits.

2. Distribution of associations

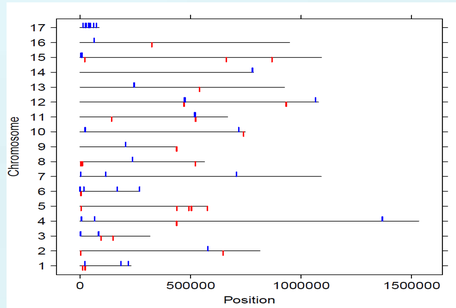


Figure 2 - Distribution of genes on chromosomal positions
The distribution of all genes related to at least one phenotype over the 16 chromosomes of *S. cerevisiae* strain S288C. Blue tags indicate a gene on the positive strand and red tags on the negative strand.

3. Dissecting enriched GO terms

Label	GO terms
Fru_A	generation of precursor metabolites and energy**; cellular respiration**; transposition$^{***}_{\bullet\bullet\bullet}$
Fru_R	generation of precursor metabolites and energy**; cellular respiration***
Fru_E	generation of precursor metabolites and energy**; cellular homeostasis*; cellular respiration**; transposition$^{***}_{\bullet}$
Man_A	cellular amino acid and derivative metabolic process*; transposition$^{***}_{\bullet\bullet\bullet}$
Man_R	generation of precursor metabolites and energy**; cellular homeostasis*
Man_E	generation of precursor metabolites and energy**; cellular respiration*; transposition$^{***}_{\bullet}$

Table 2 - Enriched Gene Ontology
Enriched Gene Ontology process terms are listed. Significance at 10 % is marked with *, 5% is marked with ** and 1 % is marked with ***. The corresponding significance based on adjusted p-values controlling the false discovery rate (q-values) are marked with $_{\bullet}$, $_{\bullet\bullet}$ and $_{\bullet\bullet\bullet}$, respectively.

4. Dissecting multivariate gene-phenotype associations

Label	Ess. genes	Paralog	Frame shifts	Stop coden	Copy no. var.
Fru_A	0.188	$2.361^{***}_{\bullet\bullet}$	0.12	1.68	$7.855^{***}_{\bullet\bullet\bullet}$
Fru_R	0.184	1.506	0.219	1.335	$5.941^{***}_{\bullet\bullet}$
Fru_E	0.216	$2.06^{**}_{\bullet}$	0.101	$2.339^{**}_{\bullet}$	$6.988^{***}_{\bullet\bullet}$
Man_A	0.335	$3.392^{***}_{\bullet\bullet\bullet}$	0.189	1.419	$3.757^{*}_{\bullet}$
Man_R	0.108	1.249	0.316	1.619	$5.178^{**}_{\bullet}$
Man_E	0.174	$2.32^{***}_{\bullet\bullet}$	0.172	$2.151^{**}_{\bullet}$	$4.055^{**}_{\bullet}$

Certain types of variations that are over-represented among the influential genes for all phenotypes. The statistics are odds-ratios indicating potential enrichment of certain gene categories among the influential genes. The categories are: Essential genes, genes with known paralogs, genes with known frame shift variation, genes with known stop codon variation and genes with known copy number variations in yeast. Significance at 10 % is marked with *, 5% is marked with ** and 1 % is marked with ***. The corresponding significance based on adjusted p-values controlling the false discovery rate (q-values) are marked with $_{\bullet}$, $_{\bullet\bullet}$ and $_{\bullet\bullet\bullet}$, respectively.

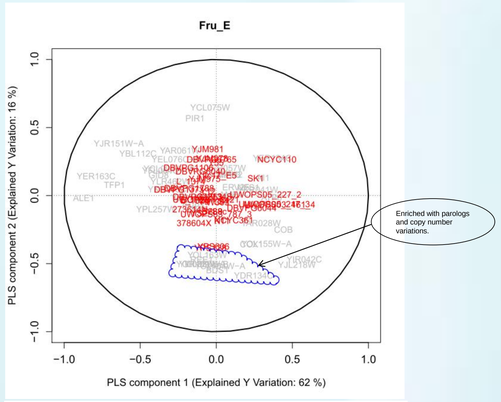


Figure 2 - Biplots for Fru_E
The biplot for Fru_E (Fructose 2% Efficiency) is presented. Genes are labeled by their names in gray color and strains are indicated by red color. For the model Fru_E most variant strains YPS606 and YPS128 are identified and their related genes are marked by the blue cloud.

Conclusion

- A multivariate approach to the analysis of the genotype-phenotype mapping based on BLAST and PLS.
- Derived results strictly adhere to the known yeast phylogeny and thus verify that the methodology extracts relevant and correct structures in the data.
- Approach is worth pursuing...

References

1. Mehmood et. al. *(2011), BMC-Bioinformatics*
2. Sæbo *et.al.* (2007), *Chemometrics and Intelligent Laboratory Systems.*

* Contact: tahir.mehmood@umb.no

2012 International Symposium on Integrative Bioinformatics
The 8th annual meeting 2nd-4th April, 2012, Hangzhou, China