Normalization and Differential Expression Analysis Method Comparison

Identifying Differentially Expressed Genes in *Saccharomyces cerevisiae*

Adriana Sedeño[1], Celia Siu[1,2], Mayumi Iwashita[2], Wendy Xu[3], Wenqiang Shi[4]

[1]Center for High-throughput Biology, [2]Department of Biochemistry, [3]Department of Statistics, [4]Centre for Molecular Medicine and Therapeutics

Background

Saccharomyces cerevisiae, oftentimes referred to as baker's yeast or brewer's yeast, is the most widely used species of yeast in the world. Industries which grew around yeast fermentation process – including biofuel[1], bakery[2], and alcoholic beverage[3] – can easily reach global market sizes of over $1 trillion today.

The growth and metabolic activities of yeast can be strongly affected by nutrient availabilities. For instance, in the industrial production of bakers' yeast, sugar-limited, aerobic cultivation at relatively low specific growth rates is essential to achieve high biomass yields (Boer et al. 2003). On the other hand, other processes such as beer fermentation occur at high concentrations of fermentable sugars and are limited by other nutrients such as oxygen and nitrogen (Boer et al. 2003). As a result, the transcriptional responses can be directly correlated to parameters such as nutritional status or stress conditions in fermentation environments (Tai et al. 2005).

In this study, the transcriptional responses, as measured by DNA microarrays, of *S.cerevisiae* sampled at four different macronutrients limitations in both the presence and absence of oxygen were reanalyzed for GSE4807 (Boer et al. 2003, Knijnenburg et al. 2007, Tai et al. 2005).

Objectives

- Identification of differentially expressed genes (or probes)
- Application and comparison of MAS5, RMA and GCRMA normalizations
- Application and comparison of SAM and Limma for DE analysis
- Enrichment analysis of the common hits found from all 3 normalization methods, Limma and SAM in 4 aerobic and anaerobic pairs.

Data

Our data was obtained from GEO[7]. The data we worked with, GSE4807 (Table 1), is a superset of GSE1723. The carbon limited samples in GSE1723 were found to be of low quality. With clustering and correlation heatmap analysis, no batch effects were observed and so the 6 (3 aerobic & 3 anaerobic) low quality carbon limited samples were replaced with the new set available in GSE4807 for our downstream analyses (Figure 1).

Before filtration, each array contained 9335 probes. The filtration step removed 60 control and 4683 probes with low variance or expressions uniformly close to background detection levels. As a result, 4637 probes were retained for DE analysis.

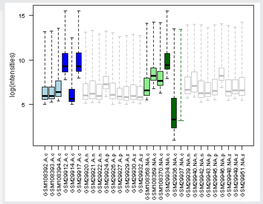


Figure 1: Boxplot of the raw data before normalization. Samples denoted by dark blue and dark green indicate the low quality samples; they replaced by the samples indicated by in the lighter colors.

Table 1: Experiment Design of GSE4807

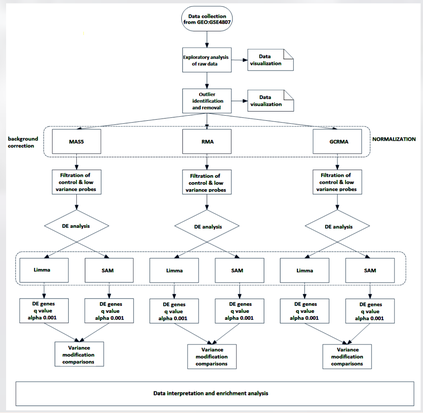
Organism: *Saccharomyces cerevisiae* (CEN.PK113-7D)

Platform: GPL90: Affymetrix Yeast Genome S98

Limitations	Aerobic	Anaerobic
Carbon	6*	6*
Nitrogen	3	3
Phosphorus	3	3
Sulfur	3	3
Total	30	

**Contains low quality samples*

Overview



Statistical Models

Normalization Methods

Table 2: Summary of RMA, MAS5 and GCRMA pre-processing methods. MAS5 was developed by Affymetrix to noramlize Affymetrix arrays.

	Background Correction	Normalization	Summarization
RMA	Signal (exponential) and noise (normal) close-form transformation	Quantile	Median Polish
MAS5	Ideal (full or partial) MM subtraction	Scaling	Tukey Biweight
GCRMA	Optical noise, probe affinity and MM adjustment	Quantile	Median Polish

DE Analysis Methods

SAM (samr package)

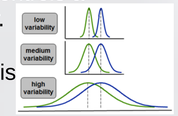
SAM was the method used in the associated publications of GSE4807. It is a non-parametric statistical technique. For our analysis, we used a moderated t-test for two-group comparison, assuming unequal variances.

The variance of SAM is adjusted as:

$$d(i) = \frac{\bar{X}_{g1} - \bar{X}_{g2}}{S_g + S_0} \quad (1)$$

,where S_g is the standard variance of each gene, and S_0 is the global variance adjustment parameter

SD of each gene and the overall distribution of the genes are estimated by permutation-based analysis. SAM performs multiple hypothesis tests, which controls a compound error rate instead of Type I error. In "samr" package, the rejection region is fixed first, then its corresponding error rate is Estimated. It is more robust against high sample variance.



Limma (Limma package)

The central idea of Limma DE analysis is to fit a linear model to the expression data for each gene. Then Empirical Bayes method will be used as a shrinkage method to borrow information across genes . This makes the analyses stable even for experiments with small sample numbers.

In Limma, the variance is adjusted as:

$$t_{gi} = \frac{\bar{X}_{g1} - \bar{X}_{g2}}{\bar{S}_g\sqrt{\frac{2}{n}}} \quad (2), \qquad \overline{S_g}^2 = \frac{d_0 {S_0}^2 + d_g {S_g}^2}{d_0 + d_g} \quad (3)$$

, where ${S_g}^2$ is the sample variance of each gene, $\overline{S_g}^2$ is the Limma standard deviation(SD), and $\bar{S}_g\sqrt{\frac{2}{n}}$ is the post Limma SD.

Since one of Limma's assumptions is equal variance across groups, we performed Levene's test. We confirmed that this assumption is valid for our dataset.

Results

Comparison of Normalization Methods Cont.

Pairwise Comparisons	SAM					limma					Overlap
aC_anC	1462	1183	1131	767		844	723	571	423		423
aN_anN	188	355	194	163		366	329	445	281		138
aP_anP	1241	806	274	233		362	298	151	133		128
aS_anS	1339	1293	803	651		676	583	508	493		447
aN_aC	1706	1380	1731	899		545	519	854	350		350
anN_anC	1308	1040	326	253		439	405	310	204		163
aC_aP	972	795	316	238		305	232	304	165		150
anC_anP	873	727	318	268		396	397	68	59		59
aC_aS	1914	1750	1023	723		760	553	888	420		405
anC_anS	1308	1040	326	253		649	544	408	308		86
aN_aS	904	692	814	534		264	267	347	198		198
anN_anS	272	265	410	206		140	135	173	106		106
aP_aN	286	138	73	59		58	60	87	44		35
anP_anN	338	242	289	149		110	86	57	42		42
aP_aS	429	239	263	184		171	126	156	102		102
anP_anS	329	260	240	174		208	165	112	92		92

SAM.GCRMA, SAM.MAS5, SAM.RMA, SAM.Overlap, limma.GCRMA, limma.MAS5, limma.RMA, limma.Overlap, Total.Overlap

Number of DE probes: 1500, 1000, 500

Table 3: Differentially expressed probes found by SAM and Limma (FDR<0.001) with GCRMA, MAS5, or RMA normalization.

* The Total.Overlap is the overlap of SAM.Overlap and limma.Overlap.

Results were found to be quite different depending on the normalization methods and DE analysis methods that are used (Table 3).

- GCRMA normalization resulted in the highest number of hits in both SAM and Limma.
- SAM reported higher number of hits than Limma.
- Most of the hits found by Limma were also found to be hits by SAM.

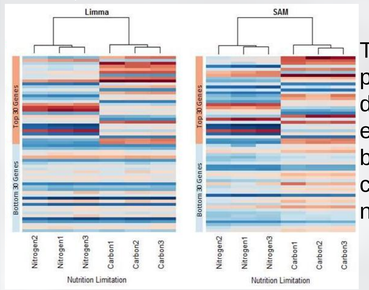


Figure 2: The top and bottom 30 DE probes using MAS5 normalizations. The q-values and absolute values of t-statistics were used to rank the probes in the SAM and Limma results respectively. Red corresponds to lower, and blue corresponds to higher levels of expressions. Anaerobic nitrogen and carbon limited samples are shown as an example.

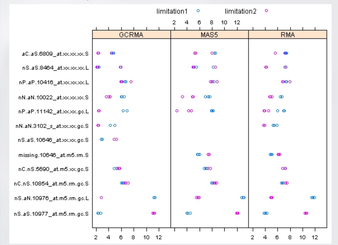


Figure 3: Visualization of hits – Comparison of RMA, MAS5, GCRMA. After a total of 32 Venn diagrams were plotted for Limma and SAM results, the top probe from each area in the venn diagrams were visualized by stripplots. Also non-DE probes were included for the comparison.

With the stripplot visualization of hits (Figure 3), we found:

- SAM can identify DE genes with higher sample variance.
- GCRMA has higher mean values
- Missing probes indicate that the filtration step did not remove the same probes.

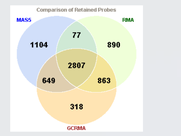


Figure 4: Probes retained after filtration. Although the same filtration was applied, the probes retained were not equal.

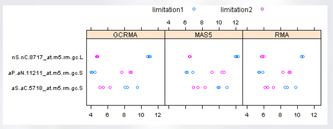


Figure 5: Visualization of hits – Comparison of Limma and SAM. We observed that SAM is not able to pick up DE genes with high variance.

Comparison of Normalization Methods

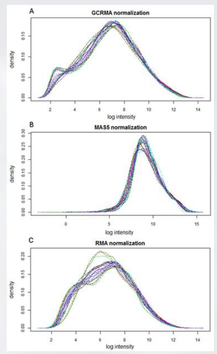


Figure 6: Distribution of expression values after (A) GCRMA, (B) MAS5 and (C) RMA pre-processing methods.

Pooled Variance Estimators (Park et al. 2003) was applied to evaluate the 3 normalization methods. We found that MAS5 has the lowest mean value and the smallest variance, while GCRMA has the highest mean value and the largest variance. According to this result, MAS5 is the best method for our dataset out of the 3 normalization methods applied. The degree of freedom, on the other hand, was observed to be the largest in MAS5.

Comparison of DE Analysis Methods

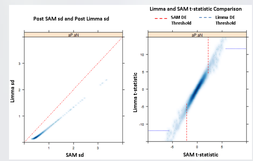


Figure 7: Comparison of post Limma and SAM standard deviations and their t-statistics. 16 of these plots were made, however, here as an example, only one is shown.

The plot of post Limma and SAM analyses of sd showed that the post SAM sd is always bigger than post Limma sd. Although this would lead to smaller t-statistics (equation 1), and thus less hits in SAM, we observed the otherwise. The reason for this resides in how these two methods estimate FDR. SAM uses resampling method to estimate the FDR rate, which is a good approximation for the real FDR and will give a low DE threshold. On the other hand, Limma uses Benjamini-Hochberg (BH) adjustment method. The BH method provides an upper bound of the real FDR and will give a high DE threshold.

Enrichment Analysis

The databse DAVID was used to perform enrichment analysis on the gene hits found in the 4 anaerobic-aerobic pairs. The list of genes were found to be fairly similar and are indeed involved in aerobic vs. anaerobic metabolism (e.g. products related to cell membrane components, mitochondria, and endoplasmic reticulum).

Discussion and Conclusions

In our dataset, MAS5 was found to be a better pre-processing method than RMA or GCRMA. We observed that SAM is more robust with large sample variances than Limma. The enrichment analysis, which was performed only for aerobic-anaerobic pairs, was concordant with the biological pathways known to be affected by the availability of oxygen. To investigate further on the missing probes in SAM hits, comparison of filtration methods maybe an interesting future topic to study.

References

1. Lane, J. "Biofuels contributed $277B to global economy: new report", "BiofuelsDigest", May 8, 2012, accessed: Mar 30, 2013, <http://www.biofuelsdigest.com/bdigest/2012/05/08/biofuels-contributed-277b-to-global-enconomy-new-report/>
2. "Conundrums in Global Bakery: A Simultaneous Quest for Health and Indulgence", "Euromonitor International", Aug 3, 2012, accessed: Mar 30, 2013, <http://blog.euromonitor.com/2012/08/conundrums-in-global-bakery-a-simultaneous-quest-for-health-and-indulgence.html>
3. MarketLine, "Alcoholic Drinks: Global Industry Guide", "Research and Markets", March 2012, accessed: Mar 30, 2013, <http://www.researchandmarkets.com/research/ea1d2d/alcoholic_drinks>
4. Boer, V. M., De Winde, J. H., Pronk, J. T., & Piper, M. D. W. (2003). The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. The Journal of biological chemistry, 278(5), 3265–74.
5. Knijnenburg, T. a, De Winde, J. H., Daran, J.-M., Daran-Lapujade, P., Pronk, J. T., Reinders, M. J. T., & Wessels, L. F. a. (2007). Exploiting combinatorial cultivation conditions to infer transcriptional regulation. BMC genomics, 8, 25.
6. Tai, S. L., Boer, V. M., Daran-Lapujade, P., Walsh, M. C., De Winde, J. H., Daran, J.-M., & Pronk, J. T. (2005). Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in Saccharomyces cerevisiae. The Journal of biological chemistry, 280(1), 437–47.
7. Barrett T. et al. (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Research, 39(Database issue):D1005-D1010.