

Topics in analysis of microarray data: differential expression, test statistics and multiple testing

Ben Bolstad

Biostatistics

University of California, Berkeley

www.stat.berkeley.edu/~bolstad

Goals of this session

- To understand and use some of the tools for analyzing pre-processed microarray data. In this session we focus on how to select differential genes.
- This session has two parts
 - Theory & Discussion of methodology
 - Hands on experimentation with BioC tools

What is differential expression?

- It is what makes the cells different
- Differential gene expression, i.e., **when**, **where**, and in **what quantity** each gene is expressed
- In comparative analysis we talk about genes being *up-regulated* or *down-regulated* when the expression level changes between two different samples or sources of mRNA.

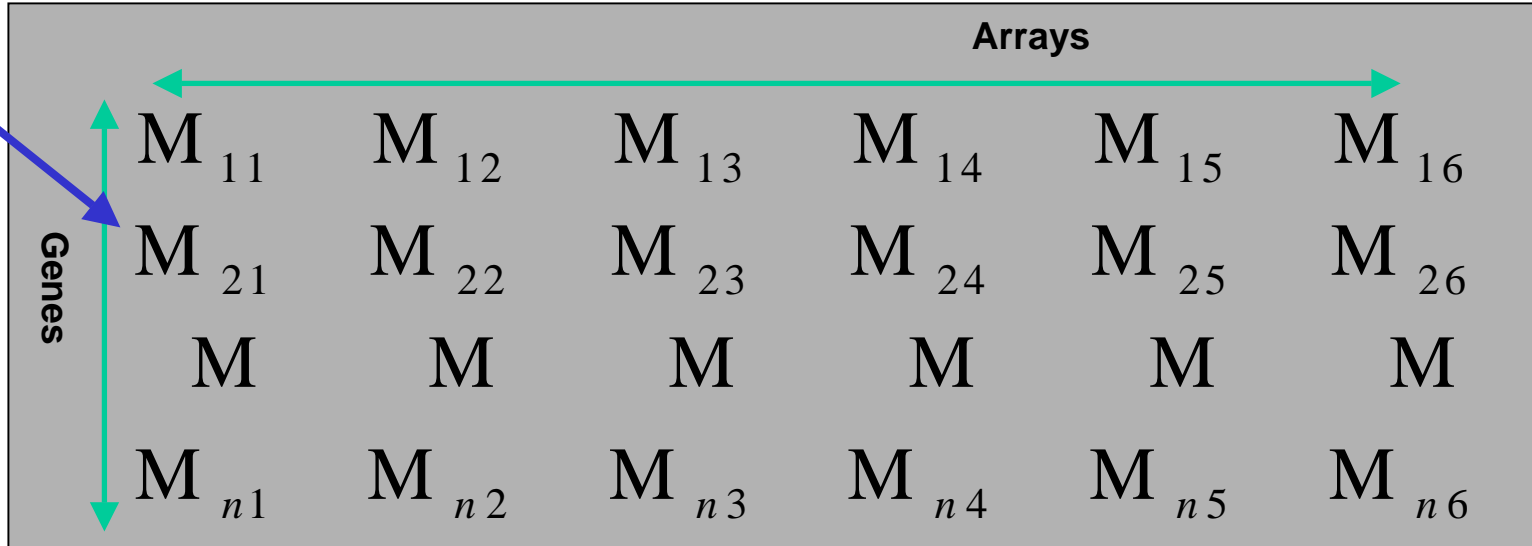
Differential expression with microarrays

- Two channel arrays (ie cDNA arrays)
 - Direct comparisons between Cy3 and Cy5 (Red/Green) channels on an array
 - Use $M = \log_2(R/G)$ as primary estimate of differential expression (this is log fold-change)
 - Multiple arrays give replication (needed for statistics)
- Single channel arrays (ie Affymetrix arrays)
 - Comparisons are between arrays. Fold-change estimates are log ratios of expression values on two arrays (or averages of two groups of arrays)

Data format

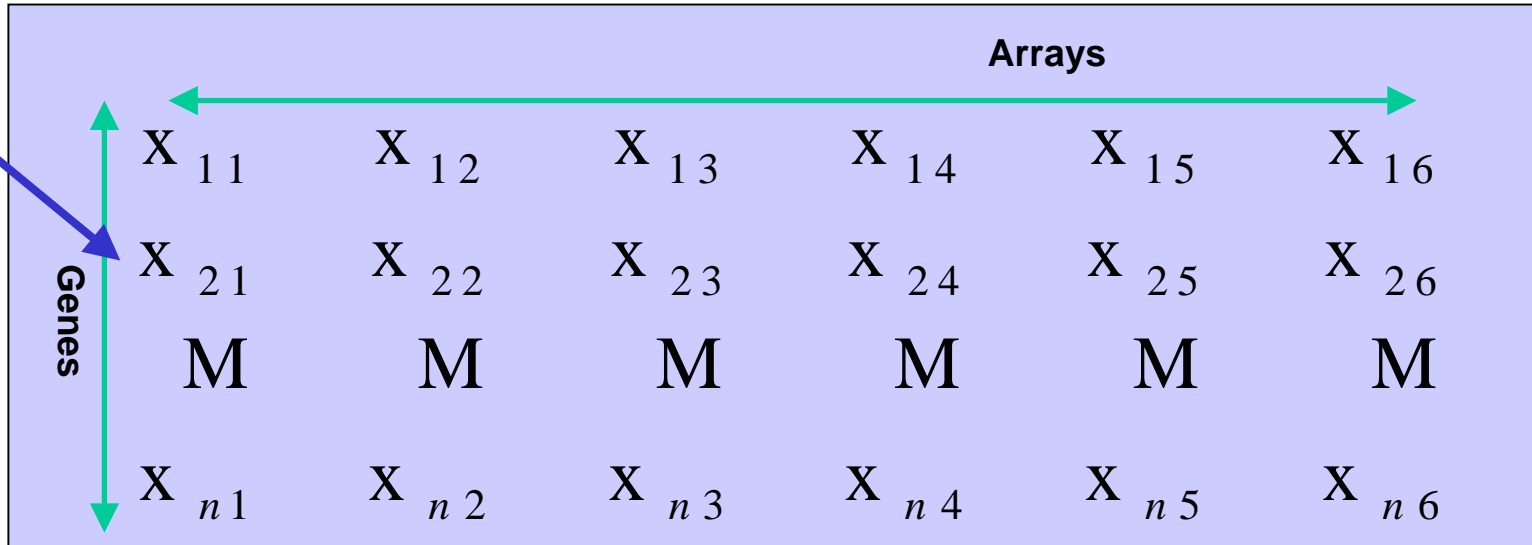
Log ratio $\log_2(R/G)$

cDNA data



Expression Value

Affymetrix data



Ranking differential expression

- Fold-change

- cDNA arrays

$$\bar{M}_i = \frac{1}{6}(M_{i1} + M_{i2} + M_{i3} + M_{i4} + M_{i5} + M_{i6})$$

- Affymetrix arrays

$$\bar{M}_i = \frac{1}{3}(x_{i1} + x_{i2} + x_{i3}) - \frac{1}{3}(x_{i4} + x_{i5} + x_{i6})$$

Ranking differential expression

- t-statistics

- cDNA arrays

$$t = \frac{\bar{M}_i}{s_i / \sqrt{n}}$$

s_i is standard deviation of M_i

- Affymetrix arrays

$$t = \frac{\bar{M}_i}{\sqrt{\frac{s_{ai}^2}{n_a} + \frac{s_{bi}^2}{n_b}}}$$

s_{ai}, s_{bi}

are standard deviations of x_i

Ranking differential expression

- Moderated t-statistic

$$t_i = \frac{\bar{M}_i}{S_i/c_i}$$

How to moderate

- Adjust variability estimate by adjusting by a prior variance

$$s_i^2 = \frac{d_g s_i^2 + d_0 s_0^2}{d_g + d_0}$$

- Prior is determined by modeling standard deviations across all genes

Moderation

- SAM
 - Tusher et al (2001)
 - Add a constant on the denominator of the test statistic
- In practice moderation increases the variability of very stable genes reducing problems of extreme t-statistics

What to do with more than two conditions?

- Work with linear models, this allows you to use all the arrays in your dataset.
- Moderated test statistics can be derived in the linear model context

How many genes?

- Two approaches
 - Try to use p-values
 - Some problems with this include that assumptions about normality might not hold, possible correlation between genes, multiple testing will require extremely small p-values, assumptions might not hold in tails
 - Take the top ranked genes and perhaps verify using other methods (eg rtPCR, Northern Blots, etc)

Multiple testing

- Consider a single test
 - Null Hypothesis (gene is not differential)
 - Alternative Hypothesis (gene is differentially expressed)
 - p value is probability of type I error (ie rejecting the null hypothesis when it is true).
 - Testing at 5% level of significance means 5% of the time will get a false positive.

Multiple testing continued ...

- Consider two tests each at 5% level. Now probability of getting a false positive is now
 - $1 - 0.95 \times 0.95 = 0.0975$
- Three tests
 - $1 - 0.95^3 = 0.1426$
- n tests
 - $1 - 0.95^n$
 - Converges towards 1 as n increases
- Small p-values don't necessarily imply significance

How do we deal with this?

- Adjust the p-values in some manner
- Two main approaches
 - Family Wise Error Rate (FWER) controlling procedures
 - FWER is probability of at least one false positive
 - False Discovery Rate (FDR) controlling procedures
 - FDR is expected value of proportion of false positives among rejected null hypotheses

FWER controlling procedures

- Bonferonni
 - $\text{adj Pvalue} = \min(n * \text{Pvalue}, 1)$
- Holm (1979)
- Hochberg (1986)
- Westfall & Young (1993) maxT and minP

FDR controlling procedures

- Benjamini & Hochberg (1995)
- Benjamini & Yekutieli (2001).

What is the difference in practice?

- FWER: gives many fewer genes (false positives), but you are likely to miss many
- FDR: if you can tolerate more false positives you will get many fewer false negatives

Some multiple testing references

- Y. Benjamini and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, Vol. 57, 289-300.
- P.H. Westfall and S.S. Young (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. *Wiley*.
- Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800-802, 1988. 1, 4
- S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65 70, 1979.1-4
- J. P. Shaffer. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561-584, 1995.

Now on to the lab