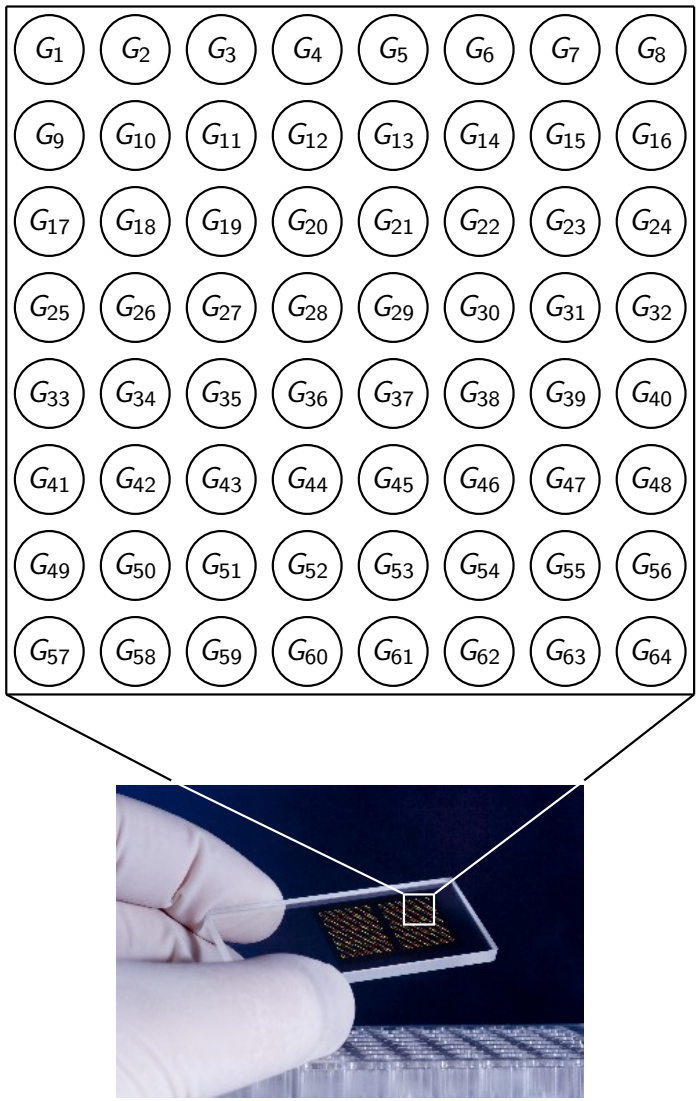
Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments

by Gordon K. Smyth
(as interpreted by Aaron J. Baraff)

STAT 572 Intro Talk

April 10, 2014

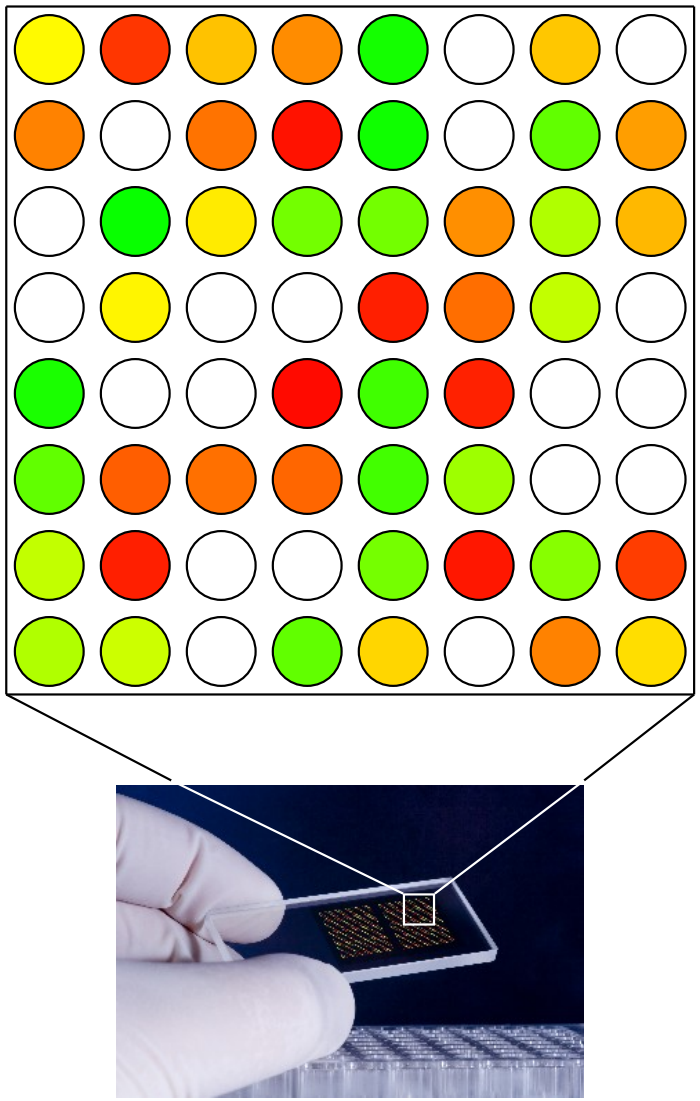
Microarray Data



Measure expression level across large numbers of genes simultaneously

- Genes express by producing mRNA $\rightarrow$ translated into proteins
- $\sim$20,000 protein-coding genes in humans
- Microarray chip contains cDNA for a different gene at each spot
- Sample cDNA hybridizes with cDNA on chip

Microarray Data



Two-color:

- cDNA from two samples dyed red and green
- Response is log-ratio of intensity

$$y_g = \log_2 \frac{R_g}{G_g}$$

- Relative expressions only (fold changes)

Single-channel:

- cDNA from a single dyed sample
- Absolute expressions

Microarray Data

Issues:

- **Expensive!** - sample sizes are low, number of genes is high

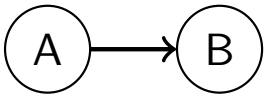
$$n << p$$

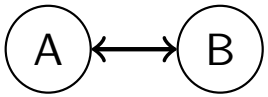
- Multiple comparisons
  - control for false discovery rate (FDR), e.g. Benjamini and Hochberg (1995, 2000)
  - often assumes independence between genes

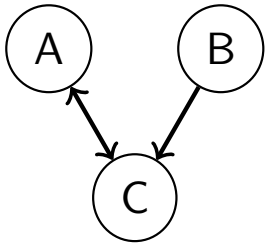
- For two-color microarrays, experimental design is more complicated

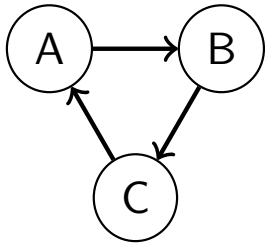
Microarray Data

Design:









Matrix:

$$X = \begin{pmatrix} 1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad X = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ -1 & -1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

Coefficients:

$$\boldsymbol{\alpha} = \begin{pmatrix} B - A \end{pmatrix} \qquad \boldsymbol{\alpha} = \begin{pmatrix} B - A \end{pmatrix} \qquad \boldsymbol{\alpha} = \begin{pmatrix} A - C \\ B - A \end{pmatrix} \qquad \boldsymbol{\alpha} = \begin{pmatrix} B - A \\ C - B \end{pmatrix}$$

Assumptions

Sample of n microarrays:

- Response vector $\mathbf{y}_g = (y_{g1}, \ldots, y_{gn})^T$ for each gene g
- Assume

$$\mathrm{E}(\mathbf{y}_g) = X\boldsymbol{\alpha}_g, \quad \text{and} \quad \mathrm{Var}(\mathbf{y}_g) = W_g\sigma_g^2$$

  for known design matrix X and weight matrix W_g
- Usually interested in contrasts of coefficients

$$\boldsymbol{\beta}_g = C^T\boldsymbol{\alpha}_g$$

  for known contrast matrix C

Assumptions

Fitting the model gives:

- Coefficient estimators $\hat{\boldsymbol{\alpha}}_g$ for $\boldsymbol{\alpha}_g$
- Contrast estimators $\hat{\boldsymbol{\beta}}_g = C^T \hat{\boldsymbol{\alpha}}_g$ for $\boldsymbol{\beta}_g$
- Variance estimators s_g^2 for σ_g^2

(Note: no assumption that $\mathbf{y}_g$ is normal or model is fit by OLS)

Assumptions

Assume:

- covariance matrices

$$\text{Var}(\hat{\boldsymbol{\alpha}}_g) = V_g \sigma_g^2 \quad \text{and} \quad \text{Var}(\hat{\boldsymbol{\beta}}_g) = C^T V_g C \sigma_g^2$$

- distributions

$$\hat{\beta}_{gi} | \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj} \sigma_g^2) \quad \text{and} \quad s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi^2_{d_g}$$

all independent, where v_{gj} is the jth diagonal element of $C^T V_g C$

Differential Expression

The problem:

- Would like to test

$$H_0 : \beta_{gj} = 0 \quad \text{vs} \quad H_1 : \beta_{gj} \neq 0$$

- **Too many genes!** - multiple comparison methods assume independence across genes
- Instead, think of p-values as statistics used to rank genes

Differential Expression

Previous methods for ranking genes:

- Fold changes - use $|\hat{\beta}_{gj}|$ directly
- t-statistics

$$|t_{gj}| = \frac{|\hat{\beta}_{gj}|}{s_g \sqrt{v_{gj}}}$$

  Problem: s_g small $\rightarrow |t_{gj}|$ large
- Offset t-statistics - inflate s_g
  - Tusher et al (2001) - minimize coefficient of variation
  - Efron et al (2001) - percentile of sample variances
- Odds ratios - Lönnstedt and Speed (2002) - empirical Bayes methods to estimate odds of differential expression - replicated experiments only
- Other methods dependent on specific designs

Differential Expression

Goal: Extend empirical Bayes method from Lönnstedt and Speed (2002) to more general experiments

Priors:

- Variance

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

- Differential expression

$$P(\beta_{gj} \neq 0) = p_j$$

- Fold change

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j} \sigma_g^2)$$

Differential Expression

After a bunch of calculgebra that I haven't done yet, we get posterior mean

$$\tilde{s}_g^2 = \frac{1}{\mathrm{E}(\sigma_g^2 | s_g^2)} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g},$$

and **moderated** t-statistic

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \sim t_{d_0 + d_g}, \quad \text{under } H_0.$$

Since this is an **empirical** Bayes method, estimate hyperparameters s_0^2 and d_0 from the data.

Results

Simulation study:

- All parameters and hyperparameters held constant except $d_0 = 1, 10, 1000$
- Moderated t has fewer false positives than other methods
- Rigging the game?

Swirl data:

- Mutation in known gene in zebrafish
- Degrees of freedom for t increases from 4 to 7.17
- Ranking more sensible than other methods

Discussion

Modern relevance:

- Method included in `R` package `limma` as part of Bioconductor
- Later papers extended the idea of sharing variance - Cui et al (2005) uses a James-Stein-type shrinkage estimator
- Applications to other *-omics* data with similar high-dimensional problems
- Digital gene expression (DGE) starting to overtake microarrays
  - observed as count data
  - modeled with overdispersed Poisson
  - empirical Bayes used to share data about overdispersion parameter across genes

Discussion

What next?

- Do the mathy stuff
  - calculation of posterior and marginal distributions
  - estimation of hyperparameters
- Data normalization
- Perform simulations
- Develop critique
  - paper makes a lot of unrealistic assumptions about distribution and independence of σ_g^2
  - imperfect method for imperfect data?

The End

Any questions?