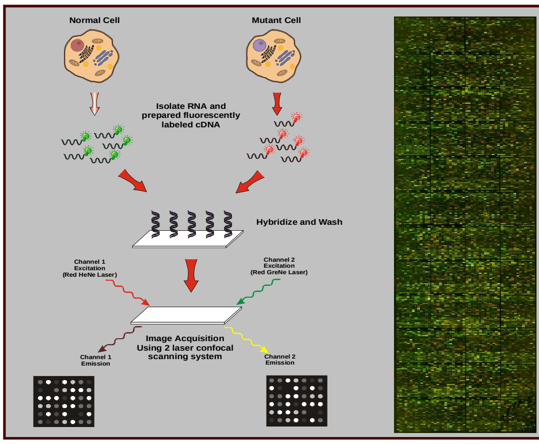
Empirical Bayes and Mixed Linear Models for Assessing Differential Expression in cDNA Microarray Experiments

Gordon Smyth

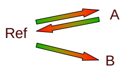
Walter and Eliza Hall Institute



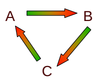
Designs → Linear Models

$$y = \log_2(R) - \log_2(G) \equiv B - A$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta \qquad \beta \equiv B - A$$



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv A - \text{Ref} \\ \beta_2 &\equiv B - A \end{aligned}$$



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \qquad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Linear Model Estimates

Obtain a linear model for each gene g

$$E(\underset{\sim}{y}_g) = X \underset{\sim}{\beta}_g \qquad \operatorname{var}(\underset{\sim}{y}_g) = W_g^{-1} \sigma_g^2$$

Estimate model by robust regression, least squares or generalized least squares to get

coefficients	$\hat{\beta}_{gj}$
standard deviations	s_g
standard errors	$\operatorname{se}(\hat{\beta}_{gj})^2 = c_{gj} s_g^2$

Parallel Inference for Genes

- 10,000-40,000 linear models
- Curse of dimensionality: Need to adjust for multiple testing, e.g., control family-wise error rate (FWE) or false discovery rate (FDR)
- Boon of parallelism: Can borrow information from one gene to another

Hierarchical Model

Normal Model	Prior
$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$	$P(\beta_{gj} \neq 0) = p$ $\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$
$s_g^2 \sim \sigma_g^2 \chi_{d_g}^2$	$\sigma_g^2 \sim s_0^2 \left(\chi_{d_0}^2 / d_0\right)^{-1}$

Reparametrization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Eliminates large t-statistics merely from very small *s*

7

Marginal Distributions

The marginal distributions of the sample variances and moderated t-statistics are mutually independent

$$s_g^2 \sim s_0^2 F_{d,d_0}$$

$$\tilde{t}_g \sim \begin{cases} t_{d_0+d} & \text{with prob } 1-p \\ \sqrt{1+c_0/c}\, t_{d_0+d} & \text{with prob } p \end{cases}$$

Degrees of freedom add!

Known result?

Estimating Prior Parameters

Marginal moments of log s² lead to estimators of s_0 and d_0:

Estimate d_0 by solving

$$\psi'(d_0/2) = \text{mean}\left\{ne_g^2 - \psi'(d_g/2)\right\}$$

where

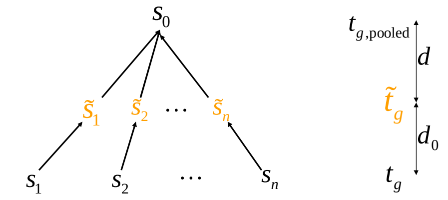
$$e_g = \log s_g^2 - \psi(d_g/2) + \log(d_g/2)$$

Finally

$$s_0^2 = \exp\left\{\bar{e} + \psi(d_0/2) - \log(d_0/2)\right\}$$

9

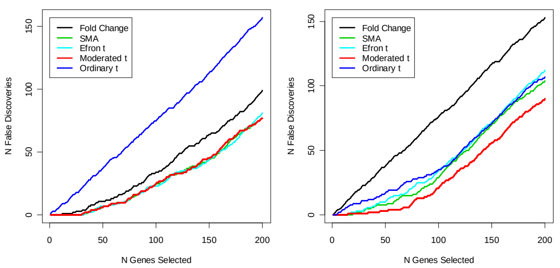
Shrinkage of Standard Deviations



The data decides whether $\tilde{t}_g$ should be closer to $t_{g,\text{pooled}}$ or to t_g

10

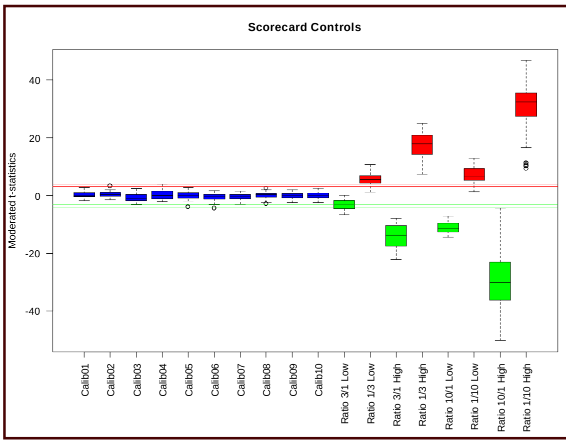
Simulations



σ^2 similar

σ^2 very different

11



Posterior Odds

Posterior probability of differential expression for any gene is

$$\frac{p(\beta \neq 0 \mid \hat\beta, s^2)}{p(\beta = 0 \mid \hat\beta, s^2)} = \frac{p}{1-p}\left(\frac{c}{c+c_0}\right)^{1/2}\left\{\frac{\tilde t^2 + d + d_0}{\tilde t^2 \frac{c}{c+c_0} + d + d_0}\right\}^{\frac{1+d+d_0}{2}}$$

Monotonic function of $\tilde t^2$ for constant *d*

Reparametrization of Lönnstedt and Speed 2002

Quantile Estimation of c_0

Let *r* be rank of $|\tilde t_g|$ in descending order, and let F(;) be the distribution function of the t-distribution. Can estimate c_0 by equating empirical to theoretical quantiles:

$$2\left[pF\left(-\sqrt{\frac{c_g}{c_g+c_0}}\,|\tilde t_g|; d_0+d_g\right) + (1-p)F(-|\tilde t_g|; d_0+d_g)\right] = \frac{r-0.5}{n}$$

Get overall estimator of c_0 by averaging the individual estimators from the top *p/2* proportion of the $|\tilde t_g|$

Duplicate spots

- Replicate spots of each gene on same array, assume duplicates at regular spacing
- Assume spatial component of correlation between duplicates is same for each gene
- Estimate spatial correlation from **consensus** estimator across genes

15

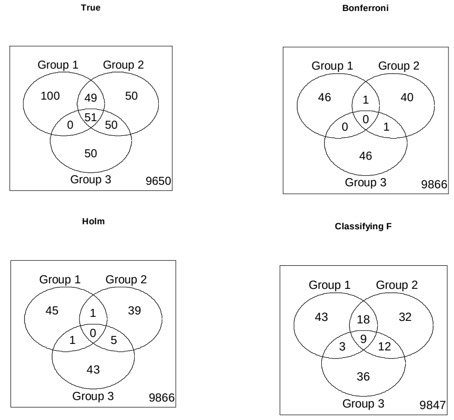
Posterior F-tests

If $\beta_g = 0$

then $\frac{\hat\beta_g^T X^T W X \hat\beta_g}{\tilde s_g^2} \sim F_{k,d+d_0}$

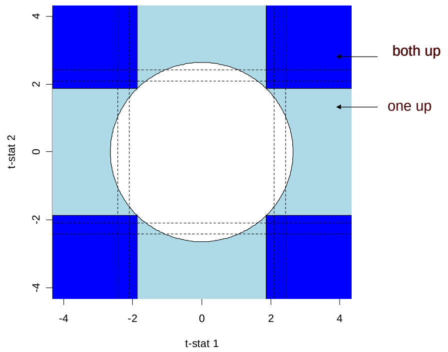
Non-null prior on β doesn't enter

16



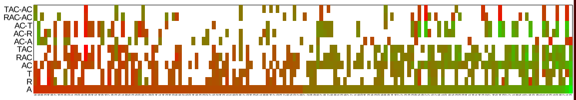
17

F-Tests as Classification Problem



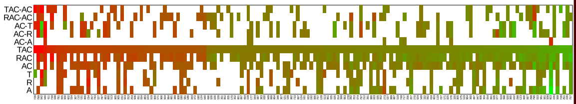
18

Stemmed Heat Diagrams



A: AML1
C: CBFb
R: ERG
T: ETS2

Top genes for AML1



Top genes for ETS2/AML1/CBFb

Acknowledgements

WEHI Bioinformatics
- Terry Speed
- Ingrid Lönnstedt
- Matt Ritchie

WEHI Scott Lab
- Joelle Michaud
- Hamish Scott

AGRF
- Steve Wilcox
- Cathy Jensen