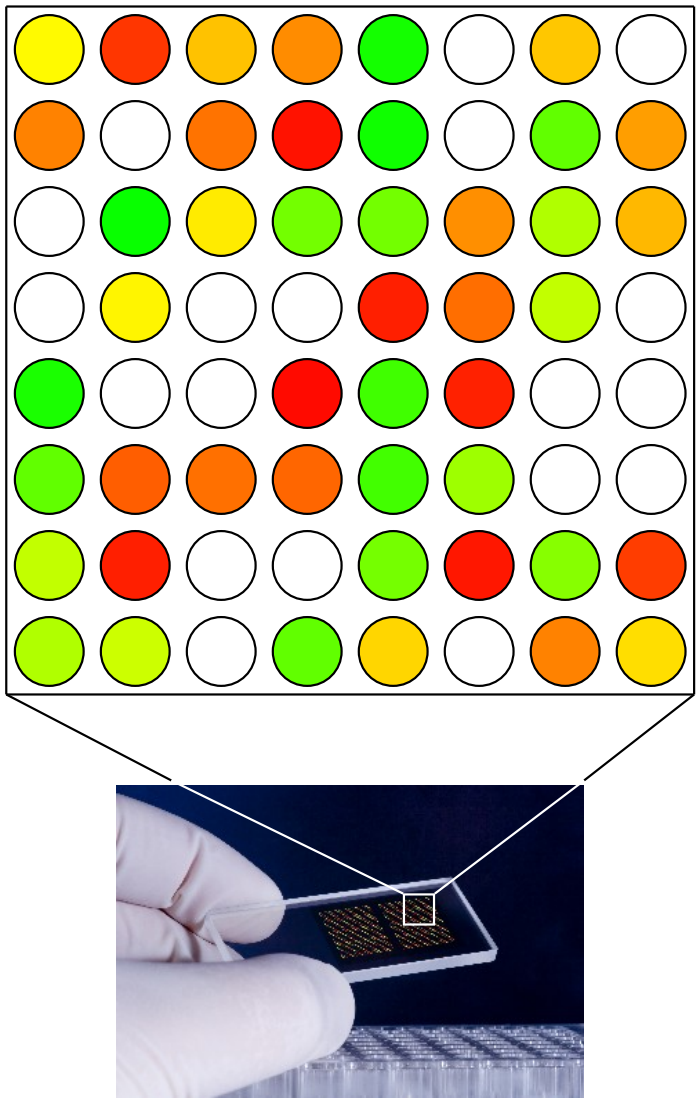
Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments

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

STAT 572 Update Talk

April 29, 2014

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

Assumptions

Sample of n microarrays:

- Response vector $\mathbf{y}_g = (y_{g1}, \ldots, y_{gn})^T$ for each gene $g = 1, \ldots, G$
- Assume the linear model

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

  for known design matrix X and weight matrix W_g
- Assume estimates have distributions

$$\hat{\boldsymbol{\beta}}_g|\boldsymbol{\beta}_g, \sigma_g^2 \sim N(\boldsymbol{\beta}_g, V_g\sigma_g^2) \quad \text{and} \quad s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$$

  independent across all genes

Differential Expression

Under $H_0 : \beta_{gj} = 0$, we have

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

Problem: Since n is often low, test statistics have high variance, leading to many false positives

Solution: Share variance information across all genes to improve estimates for σ_g^2

Bayesian Estimation of σ_g^2

Assume prior distribution on σ_g^{-2}:

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

with hyperparameters s_0^2 and d_0

Through conjugacy, we get posterior distribution:

$$\sigma_g^{-2} | s_g^2 \sim \frac{1}{d_g s_g^2 + d_0 s_0^2} \chi^2_{d_g + d_0}$$

Bayesian Estimation of σ_g^2

Now estimate σ_g^2 using the posterior mean

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

From this, we get the **moderated** t-statistic

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

(Note: $\tilde{t} \to t$ as $d_0 \to 0$, and $\tilde{t} \to c\hat{\beta}$ as $d_0 \to \infty$)

Marginal Distributions

Under $H_0 : \beta_{gj} = 0$, we have

$$\begin{aligned} p(\tilde{t}_{gj}, s_g^2 | \beta_{gj} = 0) &= \tilde{s}_g v_g p(\hat{\beta}_{gj}, s_g^2 | \beta_{gj} = 0) \\ &= \tilde{s}_g v_g \int p(\hat{\beta}_{gj} | \sigma_g^{-2}, \beta_{gj} = 0) p(s_g^2 | \sigma_g^{-2}) \pi(\sigma_g^{-2}) d\sigma_g^{-2} \\ &= [\text{pdf for } t_{d_g + d_0}] \times [\text{pdf for } s_0^2 F_{d_g, d_0}] \end{aligned}$$

Therefore,

$$\tilde{t}_{gj} \sim t_{d_g + d_0} \quad \text{and} \quad s_g^2 \sim s_0^2 F_{d_g, d_0}$$

and they are independent

Estimation of Hyperparameters

Want to use s_g^2 across all genes to estimate s_0^2 and d_0

Let $z_g = \log s_g^2$ (Fisher's z):

- $\mathrm{E}(z_g) = \log s_0^2 + \psi(d_g/2) - \psi(d_0/2) + \log(d_0/d_g)$
- $\mathrm{Var}(z_g) = \psi'(d_g/2) + \psi'(d_0/2)$

Method of moments! Solve:

$$\psi'(d_0/2) = \frac{1}{G}\sum_{g=1}^{G}[(z_g - \bar{z})^2 - \psi'(d_g/2)]$$

$$\log s_0^2 = \frac{1}{G}\sum_{g=1}^{G}[z_g - \psi(d_g/2) + \psi(d_0/2) - \log(d_0/d_g)]$$

Simulation Study - Setup

Data sets simulated under the assumed model:

$$\begin{aligned}
\hat{\beta}_g | \beta_g, \sigma_g^2 &\sim N(\beta_g, v_g \sigma_g^2) \\
s_g^2 | \sigma_g^2 &\sim \sigma_g^2 \chi_{d_g}^2 / d_g \\
\beta_{gj} | \sigma_g^2, \beta_g \neq 0 &\sim N(0, v_0 \sigma_g^2) \\
\sigma_g^{-2} &\sim \chi_{d_0}^2 / (d_0 s_0^2)
\end{aligned}$$

Using the parameters:

- $G = 15{,}000$ (300 differentially expressed)
- $d_g = 4$, $v_g = 1/3$, $v_0 = 2$, $s_0^2 = 4$
- $d_0 = 1, 4, 1000$, more to less variable

Simulation Study - Setup

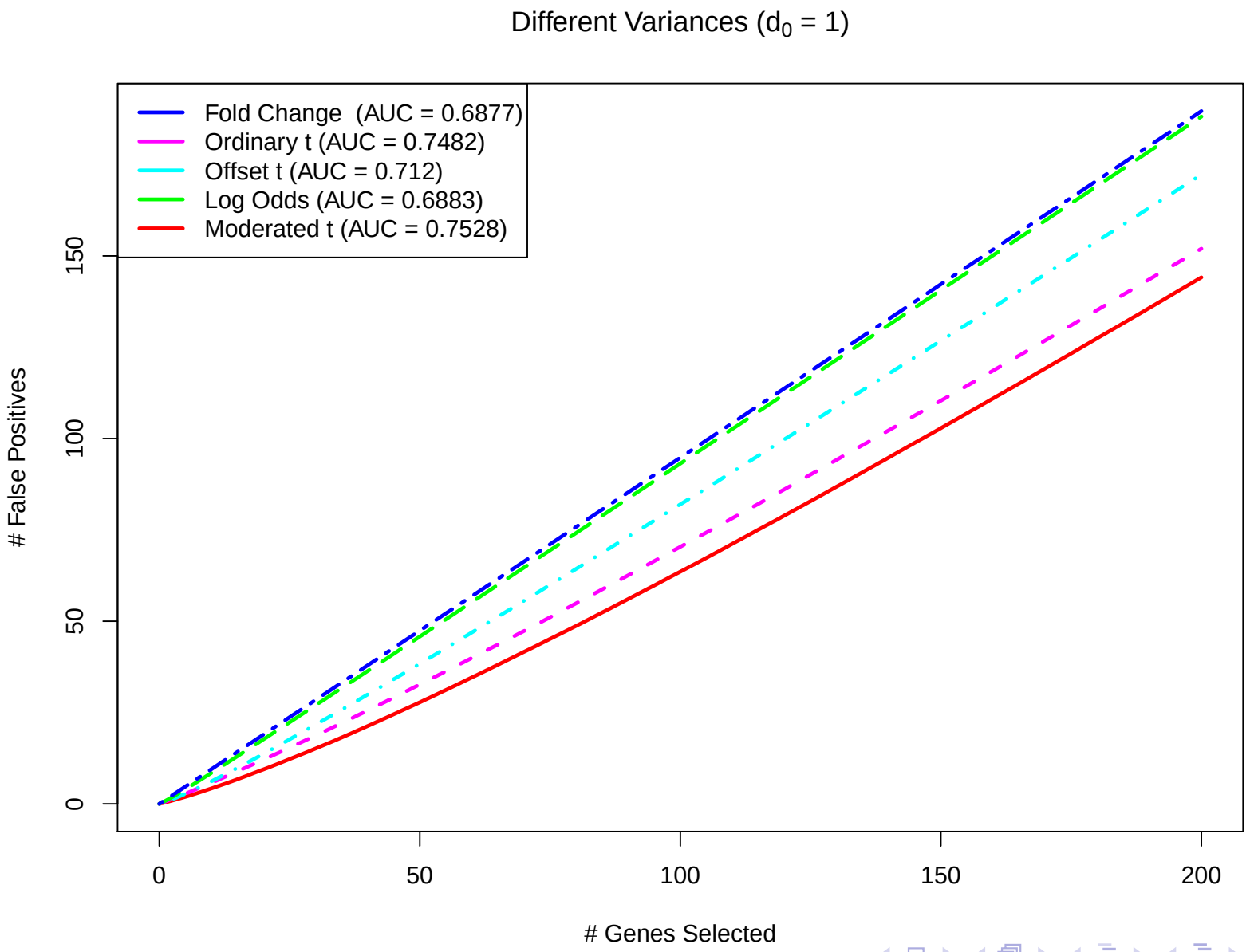
The following statistics were compared:

1. **Fold Change**: $M_g = \hat{\beta}_g$
2. **Ordinary t** - Student (1908): $t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v_g}}$
3. **Offset t** - Efron et al (2001): $t_g^* = \frac{\hat{\beta}_g}{(s_g + s_{0.9})\sqrt{v_g}}$
4. **Log Odds** - Lönnstedt and Speed (2002):

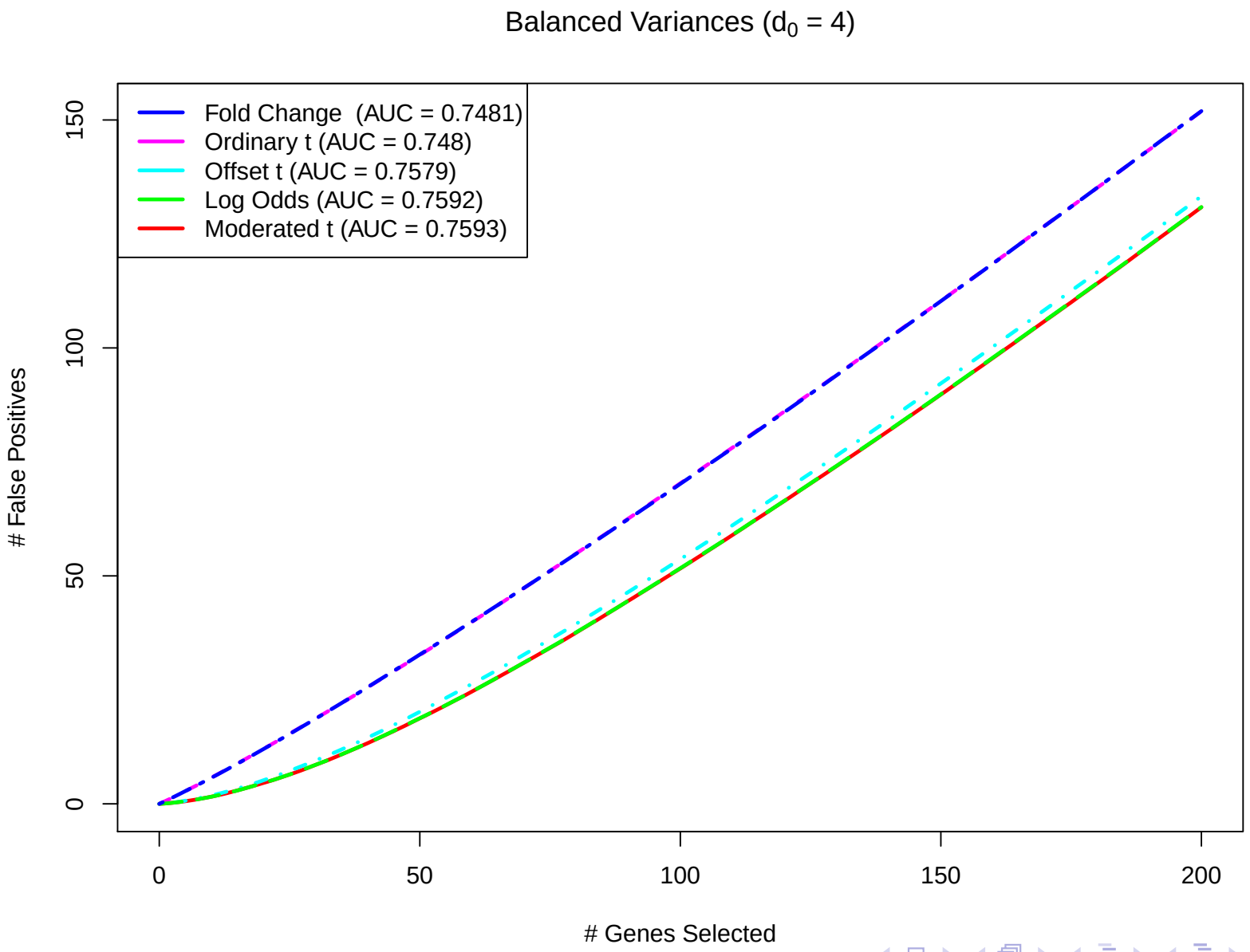
$$B_g = \log \frac{P(\beta_g \neq 0 | \hat{\beta}_g, s_1^2, \ldots, s_G^2)}{P(\beta_g = 0 | \hat{\beta}_g, s_1^2, \ldots, s_G^2)}$$

5. **Moderated t** - My paper!: $t_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}}$

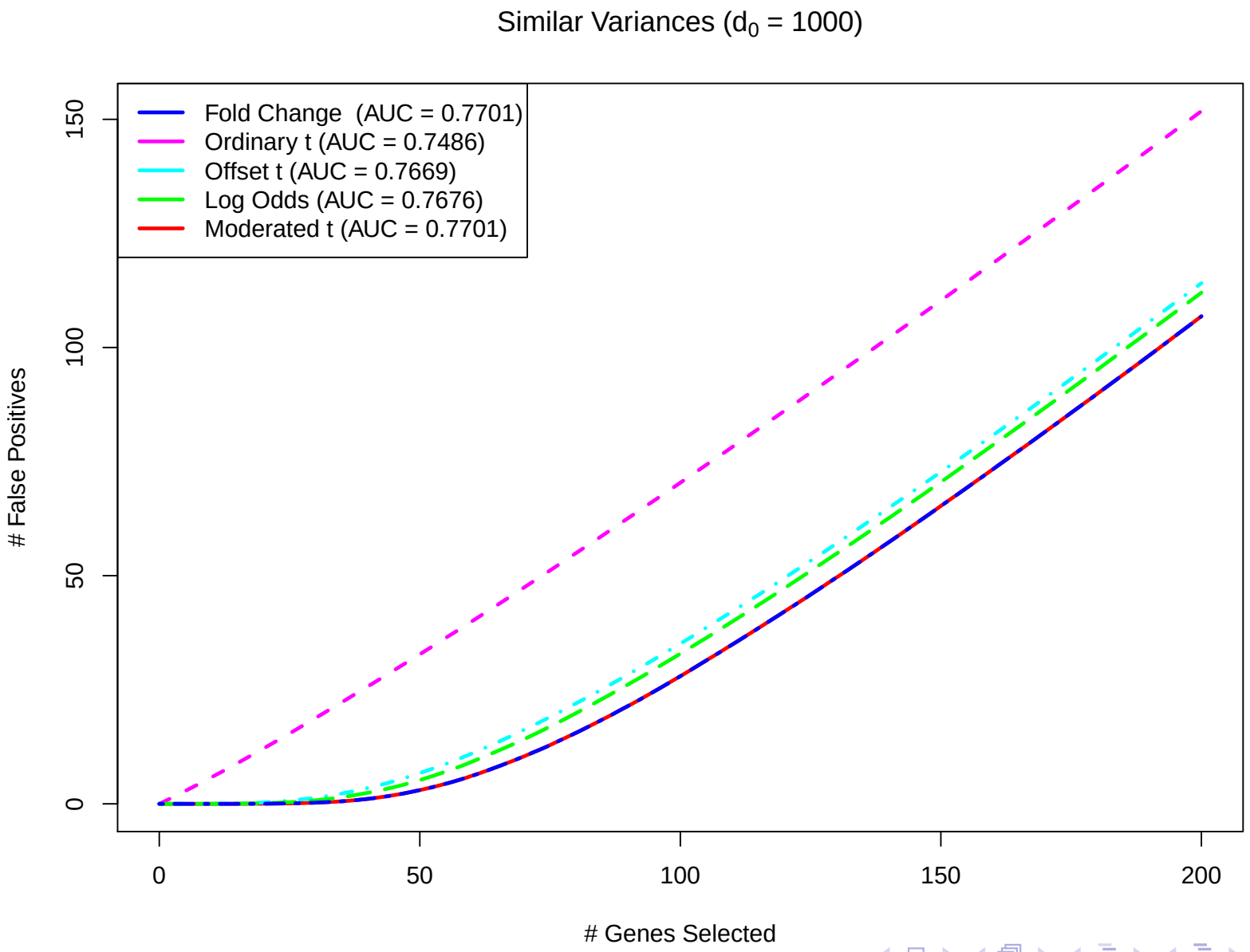
Simulation Study - Results



Simulation Study - Results



Simulation Study - Results



Simulation Study - Redux

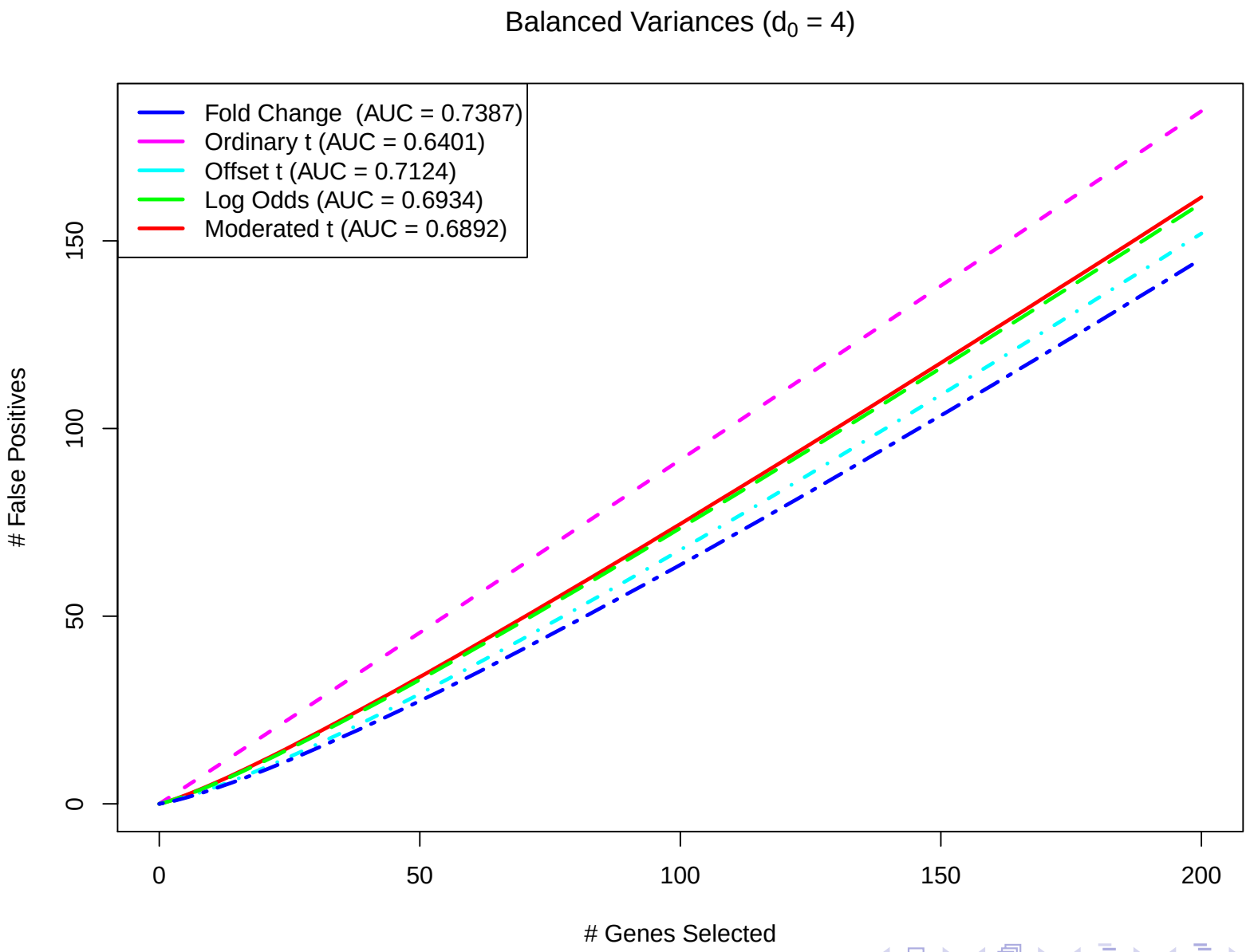
Unfair to simulate data from the assumed hierarchical model?

Everything the same except:

$$\hat{\beta}_g | \beta_g, \sigma_g^2 \sim N(\beta_g, v_g(1 + |\beta_g|)\sigma_g^2)$$
$$s_g^2 | \sigma_g^2 \sim (1 + |\beta_g|)\sigma_g^2 \chi_{d_g}^2 / d_g$$

Residual variance proportional to fold change

Simulation Study - Redux



Swirl Data

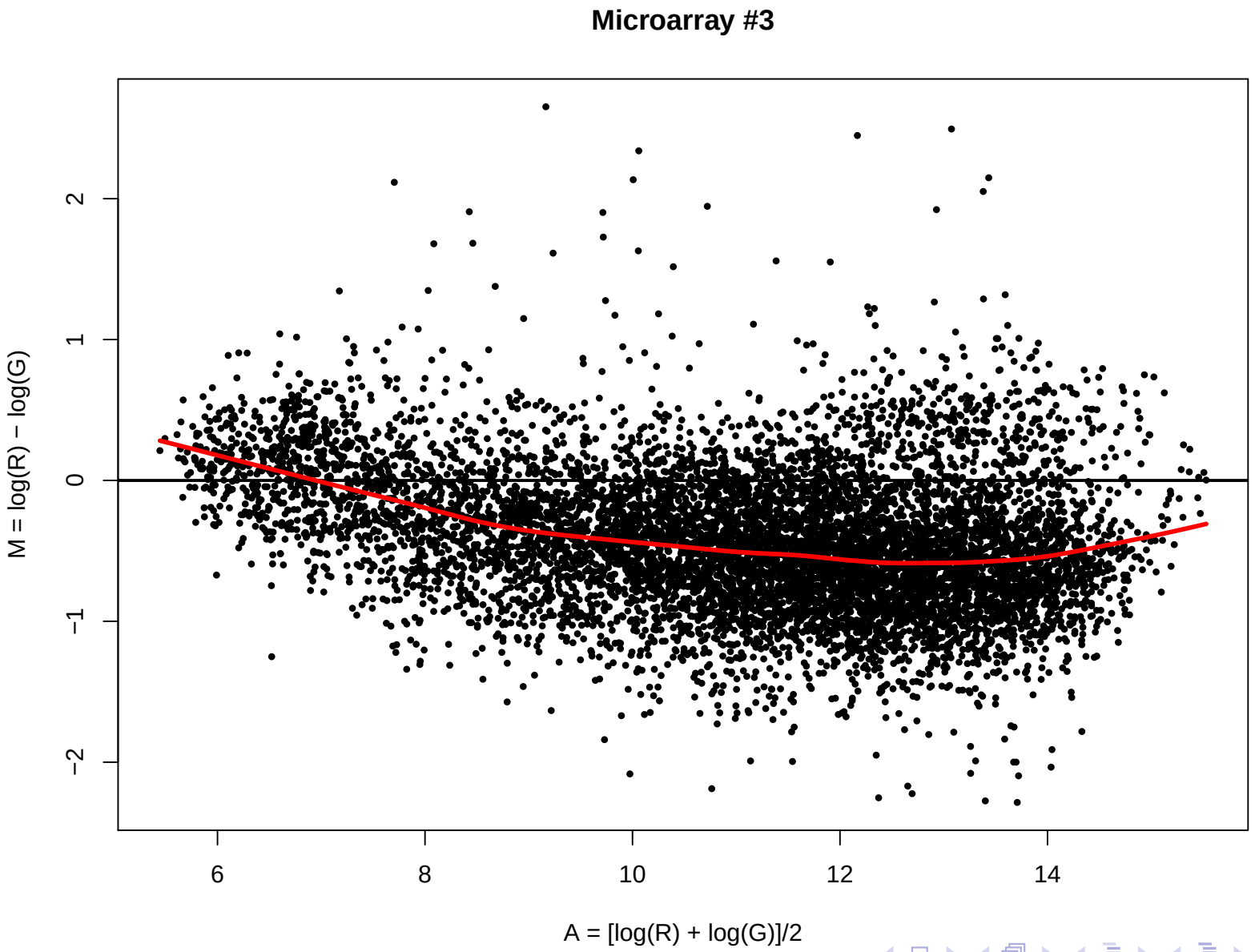
Example data:

- Swirl - mutation in BMP2 gene of zebrafish affecting dorsal/ventral body axis
- 4 microarrays - 2 dye-swap pairs

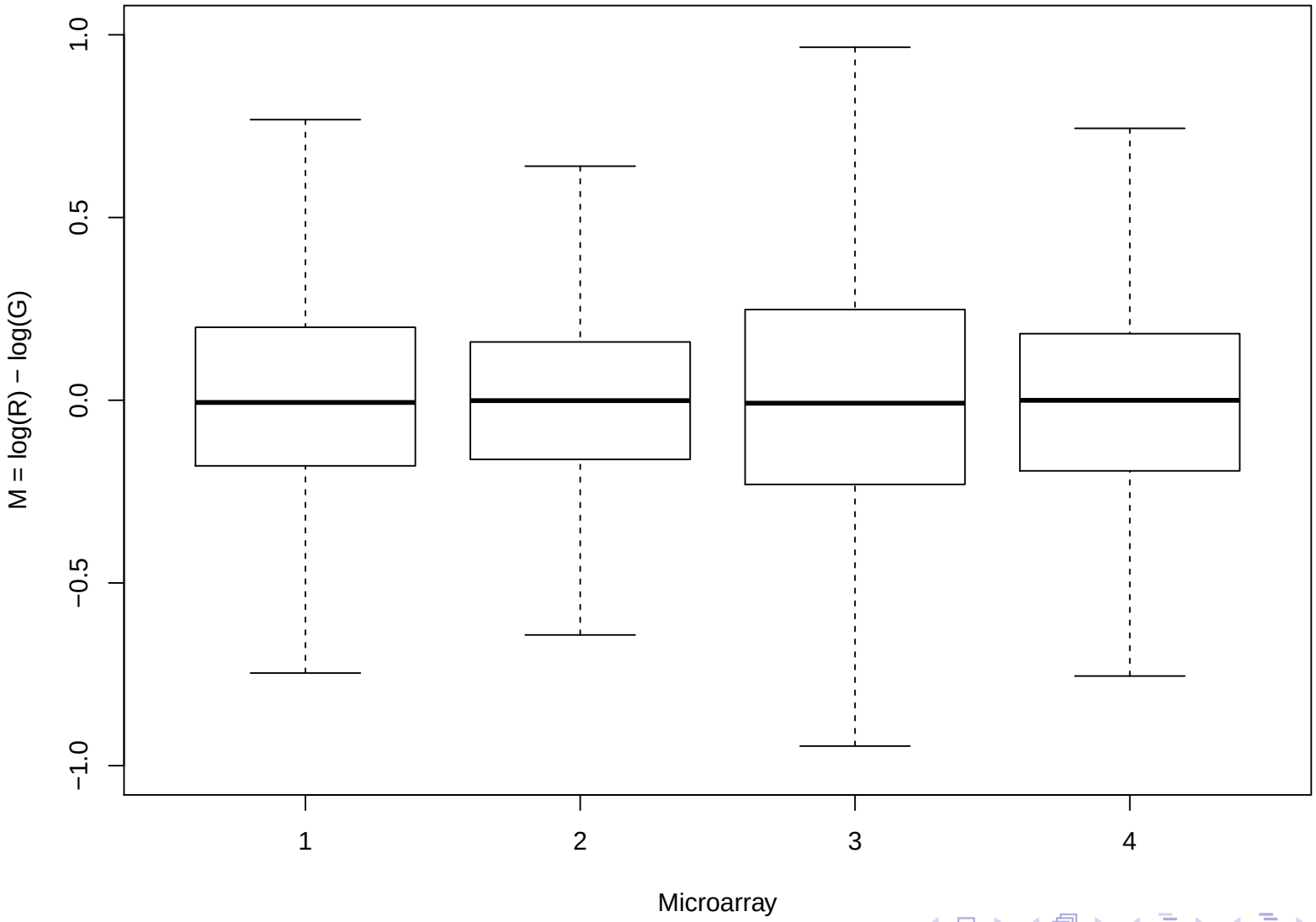
$$X = (1, -1, 1, -1)^T$$

- 8448 spots (genes) on array
- Raw data must be normalized first

Normalization - Within



Normalization - Between



Swirl Results

Prior degrees of freedom $d_0 = 4.17$

- total df increases from 3 to 7.17
- 54% reduction in variance of t under H_0

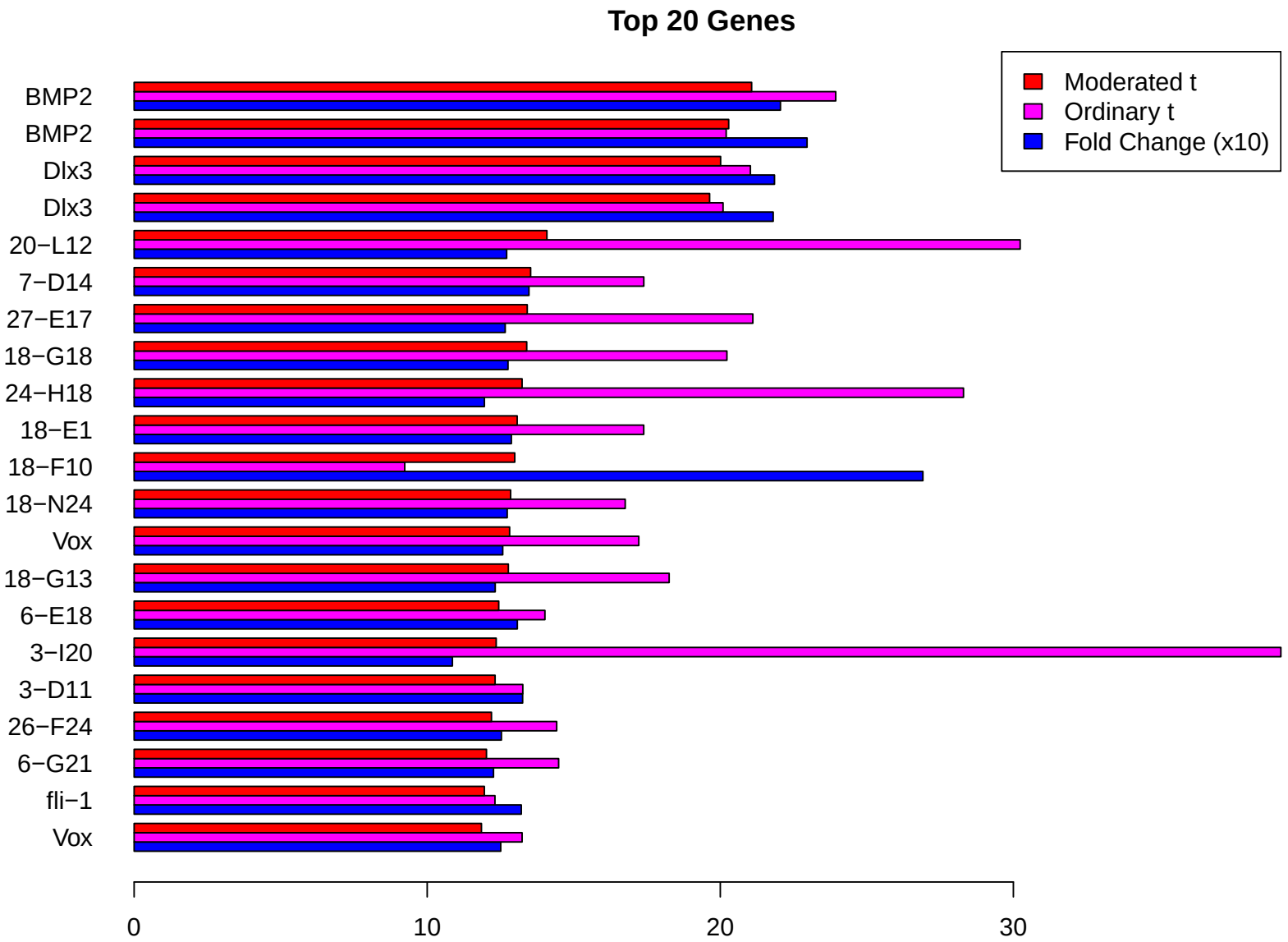
Prior variance $s_0^2 = 0.0509$

- less then mean, greater than median

Moderated t finds BMP2 and Dix3 (a known target) more clearly than other statistics

(These match the paper exactly! Yay!)

Swirl Results



Conclusions

- Need for sharing variance information across genes
- Empirical Bayes uses data to estimate hyperparameters
- Simulations showed method works well under model assumptions
  - best when variances are balanced
  - robustness issues when model is violated
- Swirl data also gives favorable results
  - small sample and balanced variances
- Classification problem, not inference?

Questions

Questions for you:

- Drop microarray design?
- Drop normalization?
- Any more ideas for breaking the method?

Questions for me?