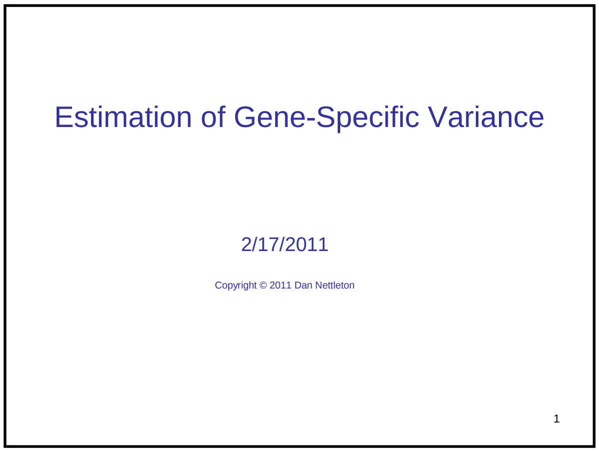
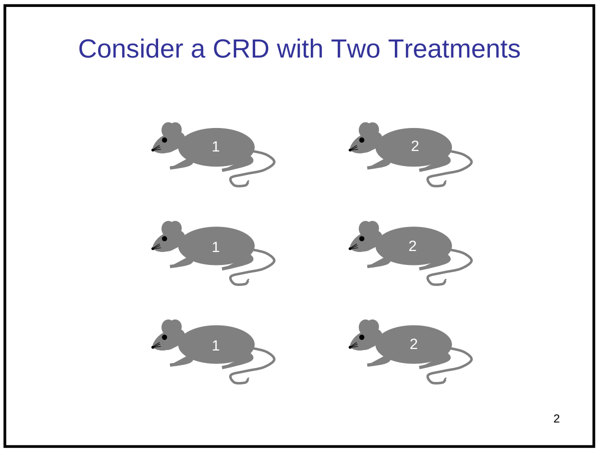
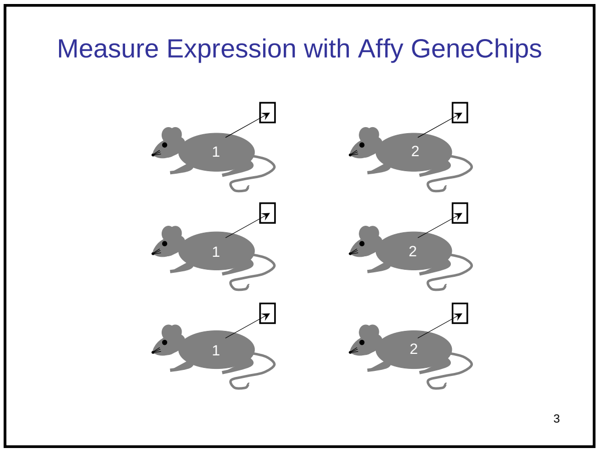


Estimation of Gene-Specific Variance

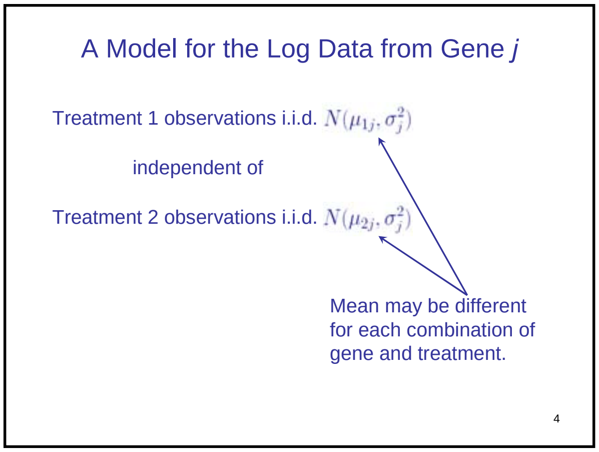
2/17/2011



Consider a CRD with Two Treatments



Measure Expression with Affy GeneChips



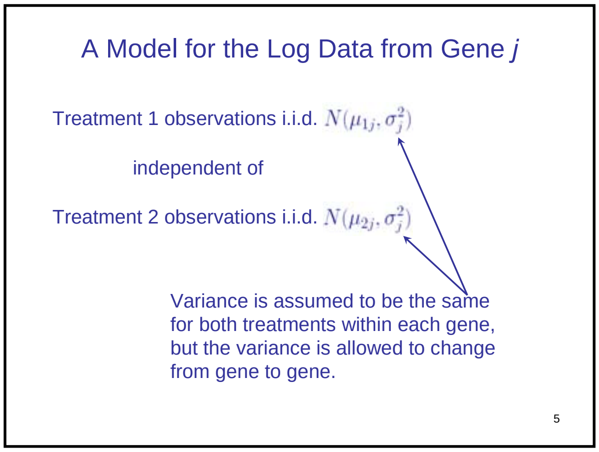
A Model for the Log Data from Gene *j*

Treatment 1 observations i.i.d. $N(\mu_{1j}, \sigma_j^2)$

independent of

Treatment 2 observations i.i.d. $N(\mu_{2j}, \sigma_j^2)$

Mean may be different for each combination of gene and treatment.



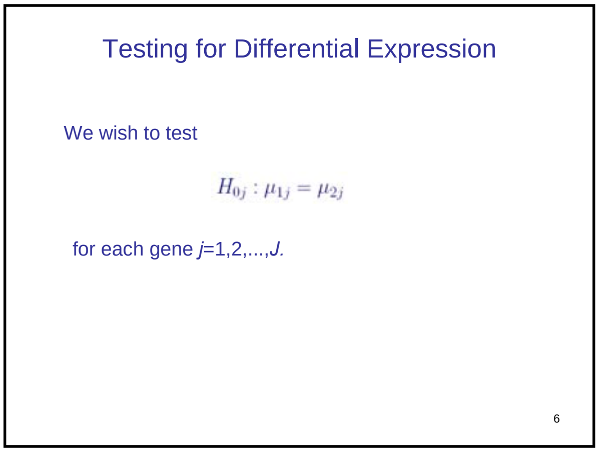
A Model for the Log Data from Gene *j*

Treatment 1 observations i.i.d. $N(\mu_{1j}, \sigma_j^2)$

independent of

Treatment 2 observations i.i.d. $N(\mu_{2j}, \sigma_j^2)$

Variance is assumed to be the same for both treatments within each gene, but the variance is allowed to change from gene to gene.



Testing for Differential Expression

We wish to test

$$H_{0j}: \mu_{1j} = \mu_{2j}$$

for each gene *j*=1,2,...,*J*.

Consider a Two-Sample *t*-Test for Each Gene

mean of treatment 1 observations for gene *j*

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{s_j^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

mean of treatment 2 observations for gene *j*

pooled variance estimate of σ_j^2

variance of trt 1 observations for gene *j*

$$s_j^2 = \frac{(n_1-1)\boxed{s_{1j}^2} + (n_2-1)\boxed{s_{2j}^2}}{(n_1-1)+(n_2-1)}$$

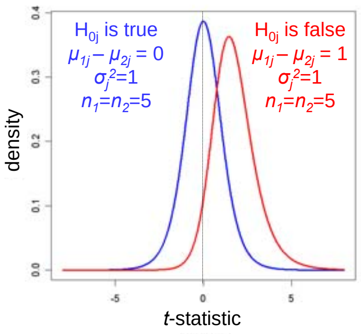
variance of trt 2 observations for gene *j*

Distribution of the *t*-Statistics under Our Model Assumptions

- Whenever H_{0j} is true, t_j will have a *t*-distribution with $d=n_1+n_2-2$ degrees of freedom.
- Whenever H_{0j} is false, t_j will have a non-central *t*-distribution with $d=n_1+n_2-2$ degrees of freedom and non-centrality parameter

$$\frac{\mu_{1j} - \mu_{2j}}{\sqrt{\sigma_j^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Distributions of *t* under null and alternative



Potential Problems with the *t*-Tests When There Are Few Degrees of Freedom per Gene

- Variance estimates based on few degrees of freedom can be unreliable.
- This can be particularly problematic if our model for the data is not quite right.
- Variances that are severely underestimated can lead to false positives while variances that are severely overestimated can lead to a loss of power for detecting differentially expressed genes.

Variance Constant across All Genes?

- Early microarray papers often assumed that variance was constant across all genes.

$$\sigma^2 \equiv \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2$$

- If this assumption holds, all the gene specific estimates can be averaged to produce a common estimate of variance for all genes.

$$s^2 = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

- Such an estimate would have $J(n_1+n_2-2)$ degrees of freedom if we were to assume all genes were independent.

Variance Constant across All Genes?

- The t-statistics could be computed as

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

and compared to a standard normal distribution to obtain a *p*-value.

- However, examination of multiple datasets suggests that this assumption is seldom tenable.

Variance Constant across All Genes?

- When the assumption is violated, false positives will tend to occur for genes with true variances larger than average, and false negatives will occur for genes with true variances smaller than average.
- Transformations have been suggested for stabilizing (making approximately constant) the variance across genes. (See next slide for some references.)
- However, my experience suggests that these transformations do not completely correct the problem.

Some References on Variance-Stabilizing Transformations for Microarray Data

- Cui, X., Kerr, M.K., Churchill, G.A. (2002). Transformation for cDNA Microarray Data. *Statistical Applications in Genetics and Molecular Biology*. Vol. 2, Issue 1, Article 4.
- Durbin, B., Rocke, D. (2004). Variance Stabilizing Transformations for Two-Color Microarrays. *Bioinformatics.* **20**, 660-667.
- Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A. Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.* **18** (Suppl. 1), S96–S104.

Hierarchical Modeling/Empirical Bayes Methods

Suppose $\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2 \sim G(\theta)$

where $G(\theta)$ is a known distribution that

depends on an unknown vector of parameters θ

that can be estimated from the data.

Some Example Papers from the Literature

- Baldi, P. and Long, A. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, No. 1, Article 3.
- Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448-2455.

Smyth (2004)

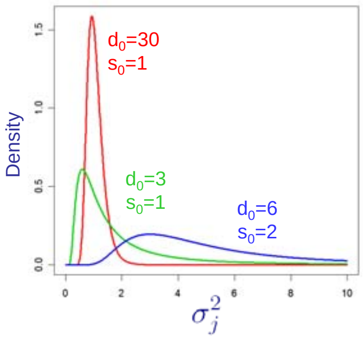
$$(s_j^2|\sigma_j^2) \sim \sigma_j^2 \frac{\chi_d^2}{d} \iff \left(\frac{ds_j^2}{\sigma_j^2}\middle|\sigma_j^2\right) \sim \chi_d^2$$

Usual assumption about s_j^2 that follows from normality and constant variance.

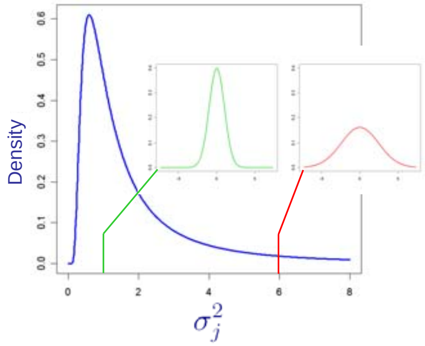
$$\left.\begin{array}{l} \sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2 \sim d_0 s_0^2 \mathrm{INV}\chi_{d_0}^2 \\ \\ \frac{d_0 s_0^2}{\sigma_1^2}, \frac{d_0 s_0^2}{\sigma_2^2}, \ldots, \frac{d_0 s_0^2}{\sigma_J^2} \sim \chi_{d_0}^2 \\ \\ \frac{s_0^2}{\sigma_1^2}, \frac{s_0^2}{\sigma_2^2}, \ldots, \frac{s_0^2}{\sigma_J^2} \sim \frac{\chi_{d_0}^2}{d_0} \end{array}\right\}$$

These are equivalent expressions that state the assumptions about the distribution of the true underlying gene-specific variances. d_0 and s_0^2 are unknown parameters to be estimated from data.

Various Probability Densities for σ_j^2



Within-Treatment Variation Varies from Gene To Gene



Smyth (2004) (continued)

Assuming independence across genes it can be shown that

$$\left(\frac{1}{\sigma_j^2}\middle|\, s_j^2, d_0, s_0^2\right) \sim \text{Gamma}\left(\frac{d+d_0}{2}, \frac{ds_j^2+d_0s_0^2}{2}\right).$$

Thus $E\left(\frac{1}{\sigma_j^2}\middle|\, s_j^2, d_0, s_0^2\right) = \frac{d+d_0}{ds_j^2+d_0s_0^2}$ and

$$E\left(\sigma_j^2\middle|\, s_j^2, d_0, s_0^2\right) = \frac{ds_j^2+d_0s_0^2}{d+d_0-2}.$$

Smyth claims $E\left(\sigma_j^2\middle|\, s_j^2, d_0, s_0^2\right) = \frac{ds_j^2+d_0s_0^2}{d+d_0}$.

Smyth's Proposed Estimator of σ_j^2

$$\tilde{s}_j^2 = \frac{ds_j^2 + d_0s_0^2}{d+d_0}$$

"Shrinks" the individual estimate s_j^2 towards s_0^2.

In practice, d_0 and s_0^2 are unknown and must be estimated from the data.

Smyth's Proposed Test Statistic

$$\tilde{t}_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{\tilde{s}_j^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Smyth refers to this as the moderated *t*-statistic.

It is the usual two-sample *t*-statistic except that s_j^2 has been replaced with $\tilde{s}_j^2$.

Distribution of the Moderated *t*-Statistic

It can be shown that $\tilde{t}_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{\tilde{s}_j^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{d+d_0}$

when H_{0j} is true and d_0 and s_0^2 are known.

Estimation of d_0 and s_0^2

It is straightforward to show that $s_j^2 \sim s_0^2 F_{d,d_0}$.

$$z_j \equiv \log s_j^2 \sim \log s_0^2 + \underbrace{\log F_{d,d_0}}_{\text{Fisher's Z Distribution}}$$

$$E(z_j) = \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d)$$

$$\text{Var}(z_j) = \psi'(d/2) + \psi'(d_0/2)$$

ψ and ψ' are the first and second derivatives of the log of the gamma function.

Estimation of d_0 and s_0^2 (continued)

$$E(z_j) = \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d)$$

$$\bar{z} \equiv \frac{1}{J}\sum_{j=1}^{J} z_j \approx \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d)$$

$$\mathrm{Var}(z_j) = \psi'(d/2) + \psi'(d_0/2)$$

$$\frac{1}{J-1}\sum_{j=1}^{J}(z_j - \bar{z})^2 \approx \psi'(d/2) + \psi'(d_0/2)$$

Estimation of d_0 and s_0^2 (continued)

$$\frac{1}{J-1}\sum_{j=1}^{J}(z_j - \bar{z})^2 \approx \psi'(d/2) + \psi'(d_0/2)$$

$$\widehat{d_0} = 2\psi'^{-1}\left(\frac{1}{J-1}\sum_{j=1}^{J}(z_j - \bar{z})^2 - \psi'(d/2)\right)$$

$$\bar{z} \equiv \frac{1}{J}\sum_{j=1}^{J} z_j \approx \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d)$$

$$\widehat{s_0^2} = \exp\{\bar{z} - \psi(d/2) + \psi(\widehat{d_0}/2) - \log(\widehat{d_0}/d)\}$$

Testing H_{0j} with the Moderated *t*-Statistic

The hope is that d_0 and s_0^2 are so well estimated because of the large number of genes that the estimates can be treated as the truth.

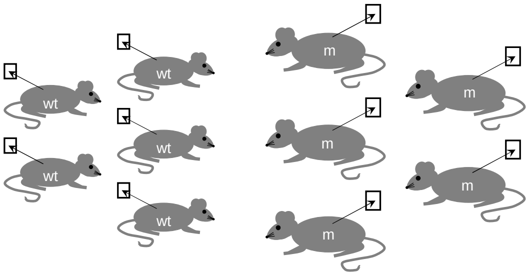
If this is reasonable, the moderated *t*-statistics with d_0 and s_0^2 replaced by their estimates can be compared to the t_{d+d_0} distribution to obtain *p*-values.

Evaluation of the Moderated *t*-Statistic

- Smyth (2004) simulate data only according to the proposed hierarchical model with complete independence across all genes.
- The performance of the moderated *t*-statistic was demonstrated to be superior to the simple two-sample *t*-statistic and other approaches with respect to ranking genes for differential expression.
- The validity of the *p*-values computed from moderated *t*-statistics was not examined.

Example: Myostatin Knockout Mice vs. Wildtype

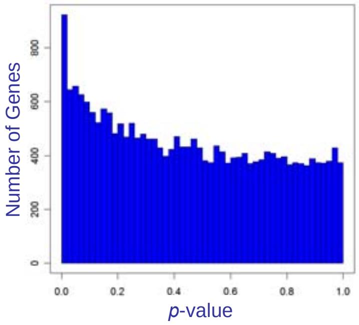
CRD with 5 Mice per Genotype



A Standard Analysis

- Two-sample *t*-tests for each gene.
- Compute *p*-values by comparing *t*-statistics to a *t*-distribution with 8 d.f.
- Convert *p*-values to *q*-values to obtain a list of differentially expressed genes and with an approximate FDR.

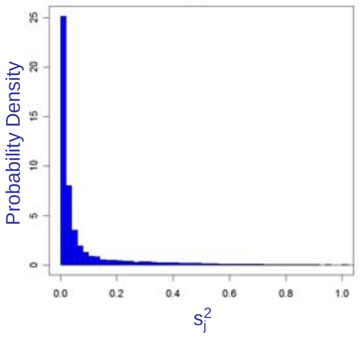
Histogram of *p*-values from the Two-Sample *t*-Tests



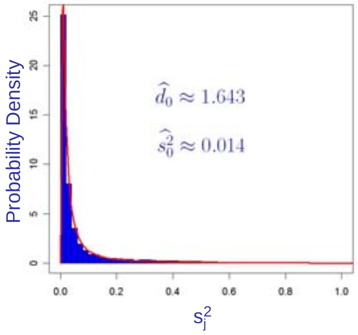
Number of Significant Genes for Various Estimated FDR Levels

FDR	Number of Genes
0.01	0
0.05	0
0.10	0
0.15	7
0.20	10
0.25	11
0.30	27
0.35	488

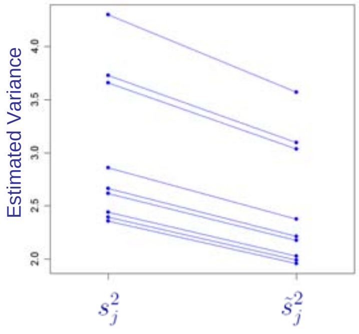
Histogram of Estimated Gene-Specific Variation

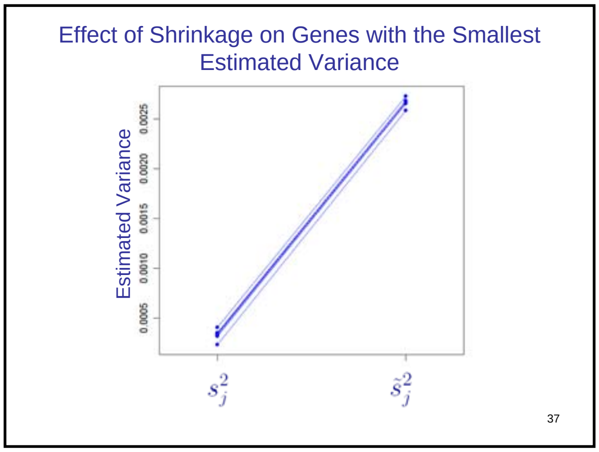


Histogram of Estimated Gene-Specific Variation with Estimated Density Based on the Model of Smyth

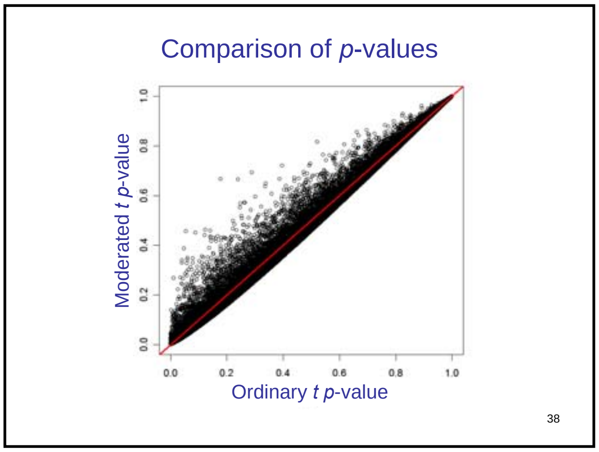


Effect of Shrinkage on Genes with the Largest Estimated Variance

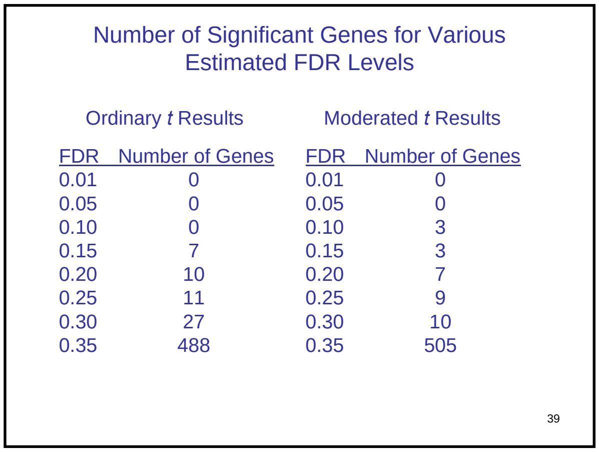




Effect of Shrinkage on Genes with the Smallest Estimated Variance



Comparison of *p*-values



Number of Significant Genes for Various Estimated FDR Levels

Ordinary *t* Results		Moderated *t* Results	
FDR	Number of Genes	FDR	Number of Genes
0.01	0	0.01	0
0.05	0	0.05	0
0.10	0	0.10	3
0.15	7	0.15	3
0.20	10	0.20	7
0.25	11	0.25	9
0.30	27	0.30	10
0.35	488	0.35	505