Selecting DE Genes: Moderated *t*-Statistic

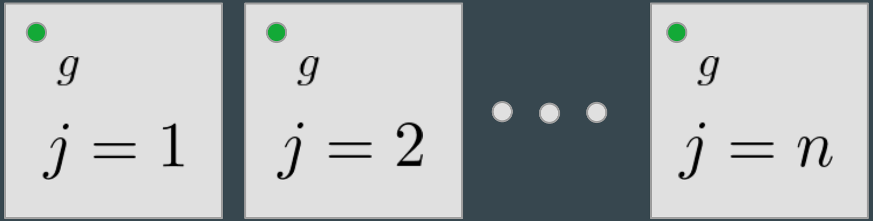
•••

by Jake Sauter

Background: *t*-Statistic

- We have previously discussed the common hypothesis test known as the *t*-Test
- The statistic for this test is

$$t = \frac{\hat{B_{gj}}}{\hat{\sigma}/\sqrt{n}}$$



g

$j = 1$

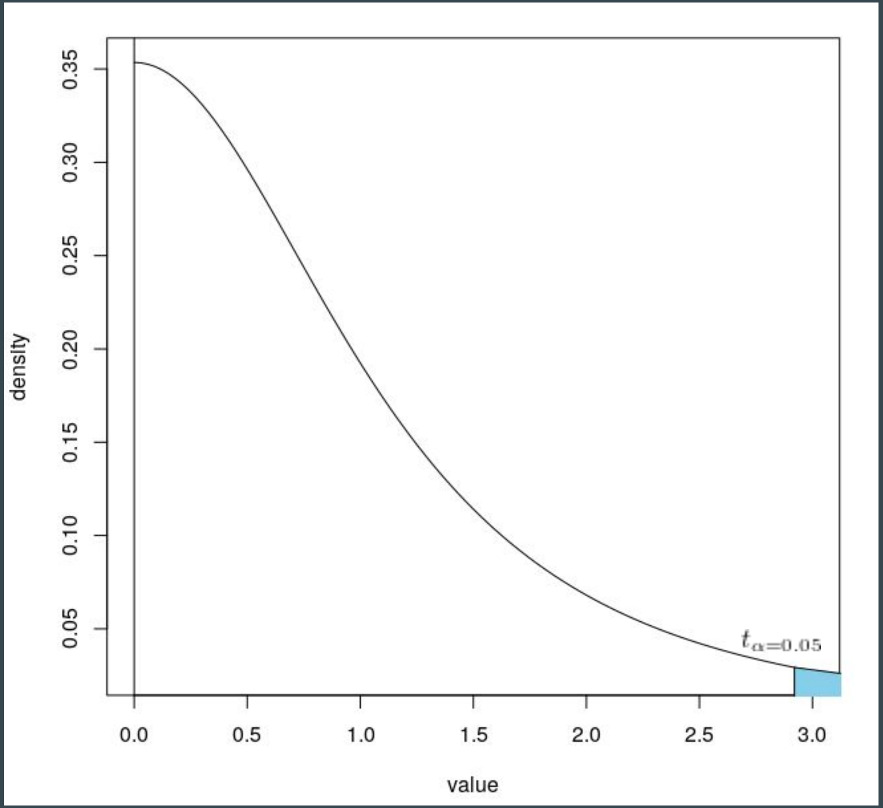
g

$j = 2$

$\cdots$

g

$j = n$



Background: Empirical Bayes Methods

- Empirical Bayes Methods are **procedures for statistical inference** in which the prior distribution is estimated from the data
- This family of methods is an approach to setting hyperparameters of known distribution to best fit the data

Background: General Linear Model

- For the General Linear Model (GLM) we assume that the observations Y_i can be modelled by a constant followed by linear scaling factors of various variables, plus an error rate for the observed sample

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

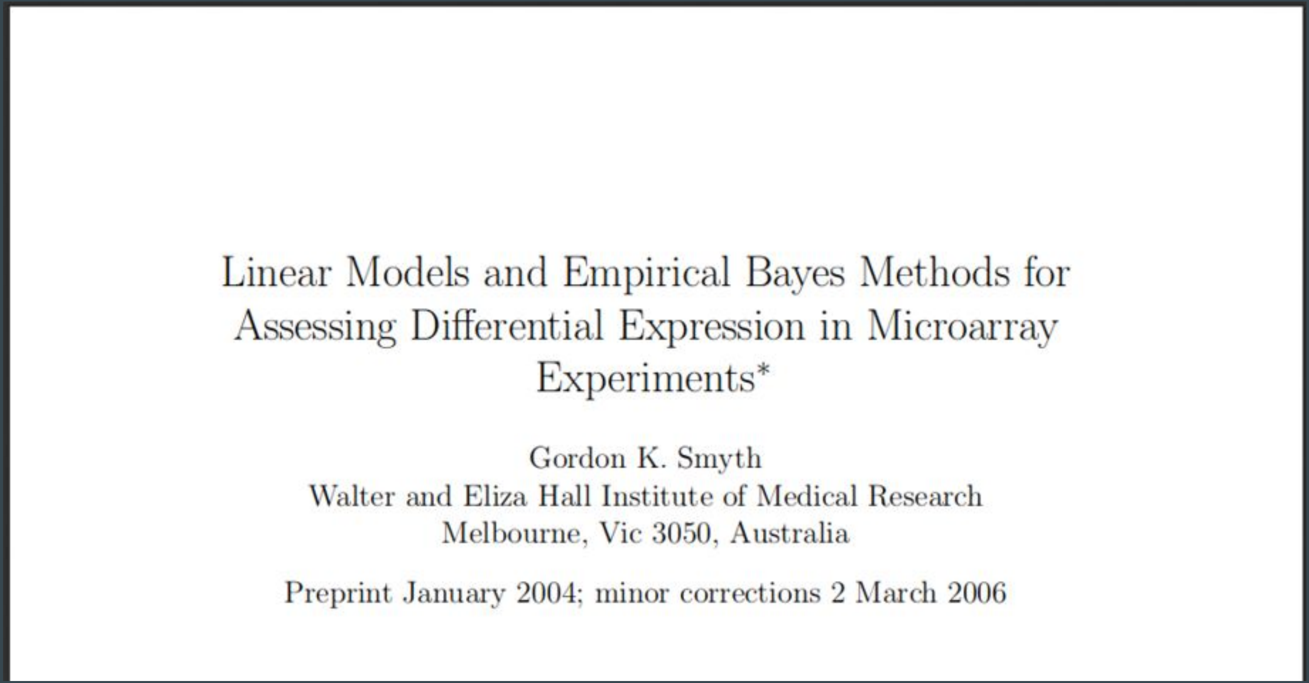
Background: Hierarchical Models

- A Hierarchical Linear Model or a Multilevel Model is statistical model of parameters that vary at more than on level
  - Generally, individual analysis and group analysis parameters share a relationship

$$
\begin{array}{lrl}
[\ \textbf{Level 1}\ ] & Y_{ij} = & \beta_{0j} + \beta X_{ij} + r_{ij} \\
 & & \downarrow \\
[\ \textbf{Level 2}\ ] & & \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + v_{0j}
\end{array}
$$

[2]

Origin of Moderated *t* Statistic



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

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

Preprint January 2004; minor corrections 2 March 2006

[1]

Linear Model Setup

- Assume that we have n microarrays with an expression vector :

$$y_g^T = (y_{g1}, \cdots, y_{g2})$$

being a vector of log-ratios or log intensities

- The general linear model is assumed as :

$$E(y_g) = X\alpha_g \text{ with } var(y_g) = W_g\sigma_g^2$$

where X is a design matrix, α_g is a coefficient vector, and W_g is a known weight matrix

Linear Model Setup

- Arbitrary contrasts of biological interest β_g can be extracted from the coefficient vector $\boldsymbol{\alpha}_g$:

$$\beta_g = C^T \alpha_g$$

- This is done with the contrast matrix C

$$C = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Linear Model Estimation

- With this linear model setup **for each gene**, the model is fit and generates estimators $\hat{\alpha_g}$ of α_g , s^2 of σ^2 , and $var(\hat{\alpha_g})$
- The estimators of contrast $\hat{\beta_g}$ and its variance estimators $var(\hat{\beta_g})$ can be derived from the model above using :

$$\beta_g = C^T \alpha_g$$

- Two assumptions about the underlying distributions is made here :
  - The contrast estimators $\hat{\beta_g}$ are normally distributed
  - The residual variances s_g^2 follow a scaled chi-square distribution

Linear Modelling

- At this point, an ordinary t statistic can be derived for the contract of interest β_{gj} being the j-th contrast for the g-th gene through the contrast estimators $\hat{\beta}_g$ and its variance estimators

- The null hypothesis $H_0 : B_{gj} = 0$ can be tested

- This process of modelling is a gene wise model fitting ignoring the parallel structure of the dependent gene expression
  - A hierarchical Bayes model can now be set up to take advantage of such information in the assessment for DE genes

Linear Modelling

- Under the assumptions that we have made about the data, the ordinary t-statistic can be calculated as :

$$t = \frac{B_{gj}}{s_g / \sqrt{v_{gj}}}$$

Hierarchical Bayes Model

- Given the large number of gene-wise linear model fits needed in a microarray experiment, it would be advantageous to make use of the parallel structure of the data
- To make use of this parallel structure, the same model is fitted to each gene
  - The key is to describe how the unknown coefficients Bgj and unknown variances σ^2 vary across genes
- In order to describe how these coefficients and variancances vary across genes, prior distributions for these sets of parameters are assumed

Hierarchical Bayes Model

- The prior information assumes that σ_g^2 is equivalent to a prior estimator s_0^2 with d_0 degrees of freedom:

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

- This describes how the variances are expected to vary across genes

Hierarchical Bayes Model

- For any given *j* , we assume that B_{gj} is non zero with known probability

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

- Where p_j is just the expected proportion of differentially expressed genes
- For those which are non zero, prior information on the coefficient is assumed equivalent to a prior observation equal to zero with unscaled variance v_{0j}

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

- This describes the expected distribution of the contrast for genes which are differentially expressed

Hierarchical Bayes Model

- Under the previously described hierarchical model, the posterior mean of $\boldsymbol{\sigma}_g^{-2}$ given s_g^2 is

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

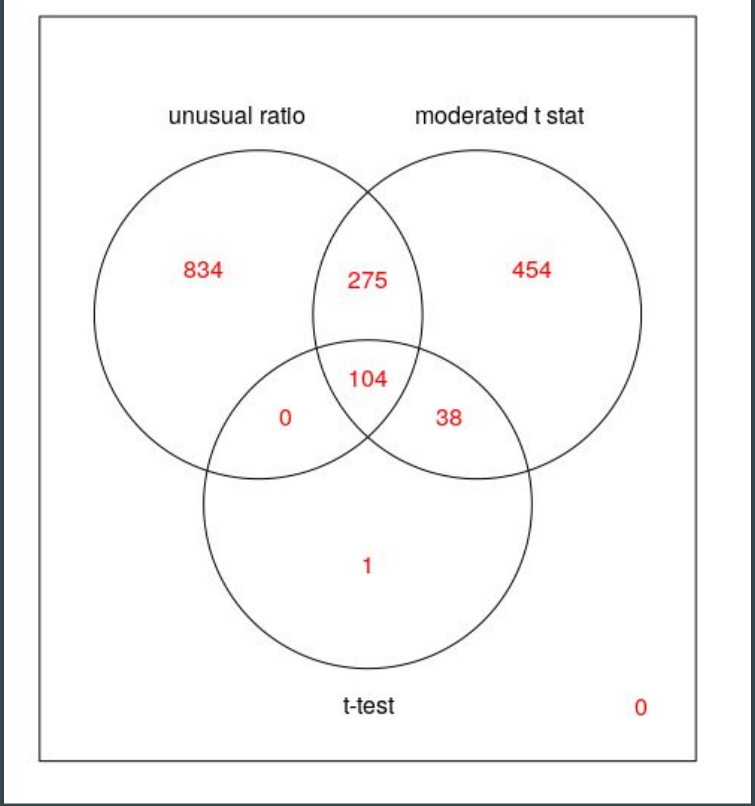
- This newly derived variance derived from our empirical bayes model can now be substituted into our t-statistic to form a more robust model

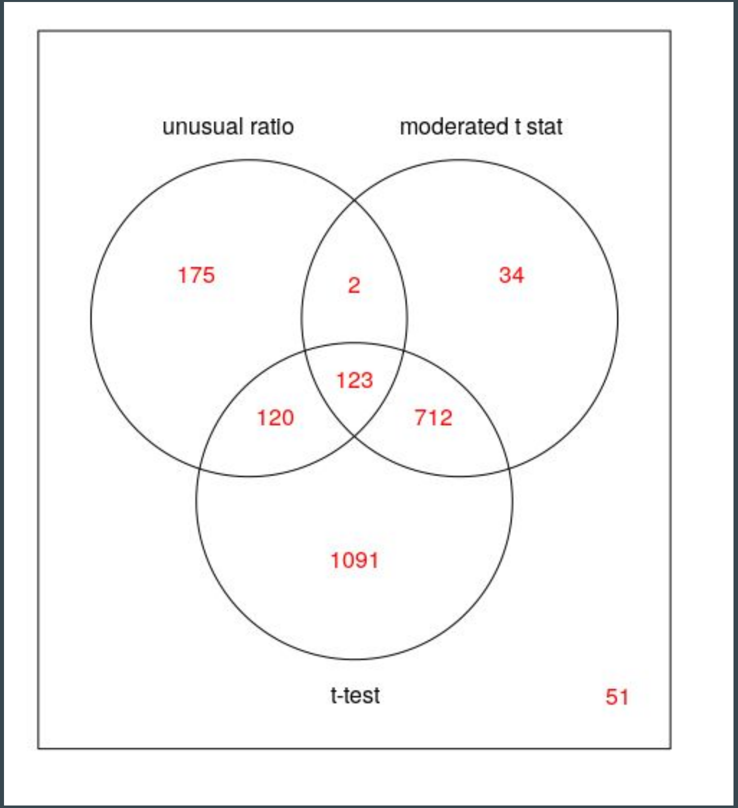
Hierarchical Bayes Model

- This model can be generated through specifying prior distributions for the unknown parameters B_{gj} and σ_g^2 in the previously described linear model
- The meta-parameters introduced in the prior distributions can be estimated from the data through an empirical Bayesian process
- The posterior residual standard deviation $\hat{s}_g^2$ can be derived from the above models, and the moderated t statistic can be defined as :

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

Results


unusual ratio
moderated t stat
834
275
454
104
0
38
1
t-test
0


unusual ratio
moderated t stat
175
2
34
123
120
712
1091
t-test
51

Discussion

- We have seen that the moderated *t* statistic approach produces more robust results than simpler non-hypothesis testing methods and hypothesis driven methods alike
- The moderated t-statistic approach is much more transparent, and better performing than SAM as it is a parametric approach with more power
- The moderated t-statistic draws power from estimating the global mean for the standard deviation and contrast parameters

References

[1] G. Smyth et al. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1027, 2004.

[2] “Introduction to Multi-Level Modeling.” YouTube, Duke, 6 Feb. 2017, www.youtube.com/watch?v=m4fx_mzlBQI.

[3] Dra˘ghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.