Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 3

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

Gordon K. Smyth*

*Walter and Eliza Hall Institute, smyth@wehi.edu.au

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress, which has been given certain exclusive rights by the author. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). http://www.bepress.com/sagmb

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

Gordon K. Smyth

Abstract

The problem of identifying differentially expressed genes in designed microarray experiments is considered. Lonnstedt and Speed (2002) derived an expression for the posterior odds of differential expression in a replicated two-color experiment using a simple hierarchical parametric model. The purpose of this paper is to develop the hierarchical model of Lonnstedt and Speed (2002) into a practical approach for general microarray experiments with arbitrary numbers of treatments and RNA samples. The model is reset in the context of general linear models with arbitrary coefficients and contrasts of interest. The approach applies equally well to both single channel and two color microarray experiments. Consistent, closed form estimators are derived for the hyperparameters in the model. The estimators proposed have robust behavior even for small numbers of arrays and allow for incomplete data arising from spot filtering or spot quality weights. The posterior odds statistic is reformulated in terms of a moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations. The empirical Bayes approach is equivalent to shrinkage of the estimated sample variances towards a pooled estimate, resulting in far more stable inference when the number of arrays is small. The use of moderated t-statistics has the advantage over the posterior odds that the number of hyperparameters which need to estimated is reduced; in particular, knowledge of the non-null prior for the fold changes are not required. The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests of composite null hypotheses through the use of moderated F-statistics. The performance of the methods is demonstrated in a simulation study. Results are presented for two publicly available data sets.

KEYWORDS: microarrays, empirical Bayes, linear models, hyperparameters, differential expression

---

*Walter and Eliza Hall Institute, 1G Royal Parade, Melbourne 3050, Australia, smyth@wehi.edu.au

Erratum

In Section 3, the posterior variance $\tilde{s}_g^2$ is introduced as the posterior mean of σ_g^2 given s_g^2. In fact, $\tilde{s}_g^{-2}$ is the posterior mean of σ_g^{-2}.

In Section 4, the second displayed equation should be

$$p(f) = \frac{a^{\nu_2/2} b^{\nu_1/2} f^{\nu_1/2-1}}{B(\nu_1/2, \nu_2/2)(a+bf)^{\nu_1/2+\nu_2/2}}.$$

In Section 6.2, all occurrences of n in equations should be G. The symbol $\bar{e}$ refers to the mean of the e_g over all G genes.

1 Introduction

Microarrays are a technology for comparing the expression profiles of genes on a genomic scale across two or more RNA samples. Recent reviews of microarray data analysis include the Nature Genetics supplement (2003), Smyth et al (2003), Parmigiani et al (2003) and Speed (2003). This paper considers the problem of identifying genes which are differentially expressed across specified conditions in designed microarray experiments. This is a massive multiple testing problem in which one or more tests are conducted for each of tens of thousands of genes. The problem is complicated by the fact that the measured expression levels are often non-normally distributed and have non-identical and dependent distributions between genes. This paper addresses particularly the fact that the variability of the expression values differs between genes.

It is well established that allowance needs to be made in the analysis of microarray experiments for the amount of multiple testing, perhaps by controlling the familywise error rate or the false discovery rate, even though this reduces the power available to detect changes in expression for individual genes (Ge et al, 2002). On the other hand, the parallel nature of the inference in microarrays allows some compensating possibilities for borrowing information from the ensemble of genes which can assist in inference about each gene individually. One way that this can be done is through the application of Bayes or empirical Bayes methods (Efron, 2001, 2003). Efron et al (2001) used a non-parametric empirical Bayes approach for the analysis of factorial data with high density oligonucleotide microarray data. This approach has much potential but can be difficult to apply in practical situations especially by less experienced practitioners. Lönnstedt and Speed (2002), considering replicated two-color microarray experiments, took instead a parametric empirical Bayes approach using a simple mixture of normal models and a conjugate prior and derived a pleasingly simple expression for the posterior odds of differential expression for each gene. The posterior odds expression has proved to be a useful means of ranking genes in terms of evidence for differential expression.

The purpose of this paper is to develop the hierarchical model of Lönnstedt and Speed (2002) into a practical approach for general microarray experiments with arbitrary numbers of treatments and RNA samples. The first step is to reset it in the context of general linear models with arbitrary coefficients and contrasts of interest. The approach applies to both single channel and two color microarrays. All of the commonly used microarray platforms such as cDNA, long-oligos and Affymetrix are therefore accommodated. The second step is to derive consistent, closed form estimators for the hyperparameters using the marginal distributions of the observed statistics. The estimators proposed here have robust behavior even for small numbers of arrays and allow for incomplete data arising from spot filtering or spot quality weights. The third step is to reformulate the posterior odds statistic in terms of a moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations. This approach makes explicit what was implicit in Lönnstedt and Speed (2002), that the hierarchical model results in a shrinkage of the gene-wise residual sample variances towards a common value, resulting in far more stable inference when the number of arrays is small. The use of moderated t-statistic has the advantage over

the posterior odds of reducing the number of hyperparameters which need to estimated under the hierarchical model; in particular, knowledge of the non-null prior for the fold changes are not required. The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests involving two or more contrasts through the use of moderated F-statistics.

The idea of using a t-statistic with a Bayesian adjusted denominator was also proposed by Baldi and Long (2001) who developed the useful cyberT program. Their work was limited though to two-sample control versus treatment designs and their model did not distinguish between differentially and non-differentially expressed genes. They also did not develop consistent estimators for the hyperparameters. The degrees of freedom associated with the prior distribution of the variances was set to a default value while the prior variance was simply equated to locally pooled sample variances.

Tusher et al (2001), Efron et al (2001) and Broberg (2003) have used t statistics with offset standard deviations. This is similar in principle to the moderated t-statistics used here but the offset t-statistics are not motivated by a model and do not have an associated distributional theory. Tusher et al (2001) estimated the offset by minimizing a coefficient of variation while Efron et al (2001) used a percentile of the distribution of sample standard deviations. Broberg (2003) considered the two sample problem and proposed a computationally intensive method of determining the offset by minimizing a combination of estimated false positive and false negative rates over a grid of significance levels and offsets. Cui and Churchill (2003) give a review of test statistics for differential expression for microarray experiments.

Newton et al (2001), Newton and Kendziorski (2003) and Kendziorski et al (2003) have considered empirical Bayes models for expression based on gamma and log-normal distributions. Other authors have used Bayesian methods for other purposes in microarray data analysis. Ibrahim et al (2002) for example propose Bayesian models with correlated priors to model gene expression and to classify between normal and tumor tissues.

Other approaches to linear models for microarray data analysis have been described by Kerr et al (2000), Jin et al (2001), Wolfinger et al (2001), Chu et al (2002), Yang and Speed (2003) and Lönnstedt et al (2003). Kerr et al (2000) propose a single linear model for an entire microarray experiment whereas in this paper a separate linear model is fitted for each gene. The single linear model approach assumes all equal variances across genes whereas the current paper is designed to accommodate different variances. Jin et al (2001) and Wolfinger et al (2001) fit separate models for each gene but model the individual channels of two color microarray data requiring the use of mixed linear models to accommodate the correlation between observations on the same spot. Chu et al (2002) propose mixed models for single channel oligonucleotide array experiments with multiple probes per gene. The methods of the current paper assume linear models with a single component of variance and so do not apply directly to the mixed model approach, although ideas similar to those used here could be developed. Yang and Speed (2003) and Lönnstedt et al (2003) take an linear modeling approach similar to that of the current paper.

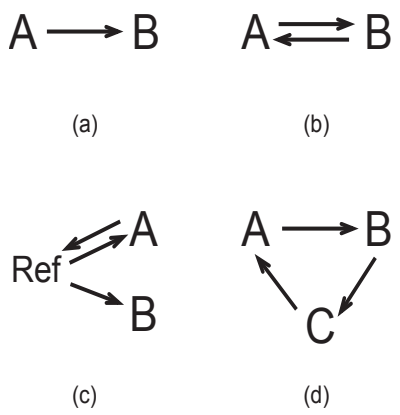

Figure 1: Example designs for two color microarrays.

The plan of this paper is as follows. Section 2 explains the linear modelling approach to the analysis of designed experiments and specifies the response model and distributional assumptions. Section 3 sets out the prior assumptions and defines the posterior variances and moderated t-statistics. Section 4 derives marginal distributions under the hierarchical model for the observed statistics. Section 5 derives the posterior odds of differential expression and relates it to the t-statistic. The inferential approach based on moderated t and F statistics is elaborated in Section 6. Section 7 derives estimators for the hyperparameters. Section 8 compares the estimators with earlier statistics in a simulation study. Section 9 illustrates the methodology on two publicly available data sets. Finally, Section 10 makes some remarks on available software.

2 Linear Models for Microarray Data

This section describes how gene-wise linear models arise from experimental designs and states the distributional assumptions about the data which will used in the remainder of the paper. The design of any microarray experiment can be represented in terms of a linear model for each gene. Figure 1 displays some examples of simple designs with two-color arrays using arrow notation as in Kerr and Churchill (2001). Each arrow represents a microarray. The arrow points towards the RNA sample which is labelled red and the sample at the base of the arrow is labelled green. The symbols A, B and C represent RNA sources to be compared. In experiment (a) there is only one microarray which compares RNA sample A and B. For this experiment one can only compute the log-ratios of expression $y_g = \log_2(R_g) - \log_2(G_g)$ where R_g and G_g are the red and green intensities for gene g. Design (b) is a dye-swap experiment leading to a very simple linear model with responses y_{g1} and y_{g2} which are log-ratios from the two microarrays and design matrix

$$X = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The regression coefficient here estimates the contrast $B - A$ on the log-scale, just as for design (a), but with two arrays there is one degree of freedom for error. Design (c) compares samples A and B indirectly through a common reference RNA sample. An appropriate design matrix for this experiment is

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

which produces a linear model in which the first coefficient estimates the difference between A and the reference sample while the second estimates the difference of interest, $B-A$. Design (d) is a simple saturated direct design comparing three samples. Different design matrices can obviously be chosen corresponding to different parametrizations. One choice is

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

so that the coefficients correspond to the differences $B - A$ and $C - B$ respectively.

Unlike two-color microarrays, single color or high density oligonucleotide microarrays usually yield a single expression value for each gene for each array, i.e., competitive hybridization and two color considerations are absent. For such microarrays, design matrices can be formed exactly as in classical linear model practice from the biological factors underlying the experimental layout.

In general we assume that we have a set of n microarrays yielding a response vector $\mathbf{y}_g^T = (y_{g1}, \ldots, y_{gn})$ for the gth gene. The responses will usually be log-ratios for two-color data or log-intensities for single channel data, although other transformations are possible. The responses are assumed to be suitably normalized to remove dye-bias and other technological artifacts; see for example Huber et al (2002) or Smyth and Speed (2003). In the case of high density oligonucleotide array, the probes are assumed to have been normalized to produce an expression summary, represented here as y_{gi}, for each gene on each array as in Li and Wong (2001) or Irizarry et al (2003). We assume that

$$E(\mathbf{y}_g) = X\boldsymbol{\alpha}_g$$

where X is a design matrix of full column rank and $\boldsymbol{\alpha}_g$ is a coefficient vector. We assume

$$\text{var}(\mathbf{y}_g) = W_g\sigma_g^2$$

where W_g is a known non-negative definite weight matrix. The vector $\mathbf{y}_g$ may contain missing values and the matrix W_g may contain diagonal weights which are zero.

Certain contrasts of the coefficients are assumed to be of biological interest and these are defined by $\boldsymbol{\beta}_g = C^T\boldsymbol{\alpha}_g$. We assume that it is of interest to test whether individual contrast values β_{gj} are equal to zero. For example, with design (d) above the experimenter might want to make all the pairwise comparisons $B - A$, $C - B$ and

$C - A$ which correspond to the contrast matrix

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

There may be more or fewer contrasts than coefficients for the linear model, although if more then the contrasts will be linearly dependent. As another example, consider a simple time course experiment with two single-channel microarrays hybridized with RNA taken at each of the times 0, 1 and 2. If the numbering of the arrays corresponds to time order then we might choose

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

so that α_{gj} represents the expression level at the jth time, and

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

so that the contrasts measure the change from time 0 to time 1 and from time 1 to time 2 respectively. For an experiment such as this it will most likely be of interest to test the hypotheses $\beta_{g1} = 0$ and $\beta_{g2} = 0$ simultaneously as well as individually because genes with $\beta_{g1} = \beta_{g2} = 0$ are those which do not change over time. Composite null hypotheses are addressed in Section 7.

We assume that the linear model is fitted to the responses for each gene to obtain coefficient estimators $\hat{\boldsymbol{\alpha}}_g$, estimators s_g^2 of σ_g^2 and estimated covariance matrices

$$\text{var}(\hat{\boldsymbol{\alpha}}_g) = V_g s_g^2$$

where V_g is a positive definite matrix not depending on s_g^2. The contrast estimators are $\hat{\boldsymbol{\beta}}_g = C^T \hat{\boldsymbol{\alpha}}_g$ with estimated covariance matrices

$$\text{var}(\hat{\boldsymbol{\beta}}_g) = C^T V_g C s_g^2.$$

The responses $\mathbf{y}_g$ are not necessarily assumed to be normal and the fitting of the linear model is not assumed to be by least squares. However the contrast estimators are assumed to be approximately normal with mean $\boldsymbol{\beta}_g$ and covariance matrix $C^T V_g C \sigma_g^2$ and the residual variances s_g^2 are assumed to follow approximately a scaled chisquare distribution. The unscaled covariance matrix V_g may depend on $\boldsymbol{\alpha}_g$, for example if robust regression is used to fit the linear model. If so, the covariance matrix is assumed

to be evaluated at $\hat{\boldsymbol{\alpha}}_g$ and is the dependence is assumed to be such that it can be ignored to a first order approximation.

Let v_{gj} be the jth diagonal element of $C^T V_g C$. The distributional assumptions made in this paper about the data can be summarized by

$$\hat{\beta}_{gj} \,|\, \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$$

and

$$s_g^2 \,|\, \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$$

where d_g is the residual degrees of freedom for the linear model for gene g. Under these assumptions the ordinary t-statistic

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

follows an approximate t-distribution on d_g degrees of freedom.

The hierarchical model defined in the next section will assume that the estimators $\hat{\boldsymbol{\beta}}_g$ and s_g^2 from different genes are independent. Although this is not necessarily a realistic assumption, the methodology which will be derived makes qualitative sense even when the genes are dependent, as they likely will be for data from actual microarray experiments.

This paper focuses on the problem of testing the null hypotheses $H_0 : \beta_{gj} = 0$ and aims to develop improved test statistics. In many gene discovery experiments for which microarrays are used the primary aim is to rank the genes in order of evidence against H_0 rather than to assign absolute p-values (Smyth et al, 2003). This is because only a limited number of genes may be followed up for further study regardless of the number which are significant. Even when the above distributional assumptions fail for a given data set it may still be that the tests statistics perform well from a ranking the point of view.

3 Hierarchical Model

Given the large number of gene-wise linear model fits arising from a microarray experiment, there is a pressing need to take advantage of the parallel structure whereby the same model is fitted to each gene. This section defines a simple hierarchical model which in effect describes this parallel structure. The key is to describe how the unknown coefficients β_{gj} and unknown variances σ_g^2 vary across genes. This is done by assuming prior distributions for these sets of parameters.

Prior information is assumed on σ_g^2 equivalent to a prior estimator s_0^2 with d_0 degrees of freedom, i.e.,

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

This describes how the variances are expected to vary across genes. For any given j, we assume that a β_{gj} is non zero with known probability

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

Then p_j is the expected proportion of truly differentially expressed genes. For those which are nonzero, prior information on the coefficient is assumed equivalent to a prior observation equal to zero with unscaled variance v_{0j}, i.e.,

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

This describes the expected distribution of log-fold changes for genes which are differentially expressed. Apart from the mixing proportion p_j, the above equations describe a standard conjugate prior for the normal distributional model assumed in the previous section. In the case of replicated single sample data, the model and prior here is a reparametrization of that proposed by Lönnstedt and Speed (2002). The parametrizations are related through $d_g = f$, $v_g = 1/n$, $d_0 = 2\nu$, $s_0^2 = a/(d_0 v_g)$ and $v_0 = c$ where f, n, ν and a are as in Lönnstedt and Speed (2002). See also Lönnstedt (2001). For the calculations in this paper the above prior details are sufficient and it is not necessary to fully specify a multivariate prior for the $\boldsymbol{\beta}_g$.

Under the above hierarchical model, the posterior mean of σ_g^2 given s_g^2 is

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

The posterior values shrink the observed variances towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom. Define the *moderated* t-statistic by

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

This statistic represents a hybrid classical/Bayes approach in which the posterior variance has been substituted into to the classical t-statistic in place of the usual sample variance. The moderated t reduces to the ordinary t-statistic if $d_0 = 0$ and at the opposite end of the spectrum is proportion to the coefficient $\hat{\beta}_{gj}$ if $d_0 = \infty$.

In the next section the moderated t-statistics $\tilde{t}_{gj}$ and residual sample variances s_g^2 are shown to be distributed independently. The moderated t is shown to follow a t-distribution under the null hypothesis $H_0 : \beta_{gj} = 0$ with degrees of freedom $d_g + d_0$. The added degrees of freedom for $\tilde{t}_{gj}$ over t_{gj} reflect the extra information which is borrowed, on the basis of the hierarchical model, from the ensemble of genes for inference about each individual gene. Note that this distributional result assumes d_0 and s_0^2 to be given values. In practice these values need to be estimated from the data as described in Section 6.

4 Marginal Distributions

Section 2 states the distributions of the sufficient statistics $\hat{\beta}_{bj}$ and s_g^2 conditional on the unknown parameters β_{gj} and σ_g^2. Here we derive the unconditional joint distribution of these statistics under the hierarchical model defined in the previous section. The unconditional distributions are functions of the hyperparameters d_0 and s_0^2. It turns out that $\hat{\beta}_{gj}$ and s_g^2 are no longer independent under the unconditional distribution but that $\tilde{t}_{gj}$ and s_g^2 are. For the remainder of this section we will assume that the calculations are being done for given values of g and j and for notational simplicity the subscripts g and j will be suppressed in $\tilde{t}_{gj}$, $\tilde{s}_g$, s_g, β_{gj}, v_{gj}, v_{0j} and p_j.

It is convenient to note first the density functions for the scaled t and F distributions. If T is distributed as $(a/b)^{1/2}Z/U$ where $Z \sim N(0,1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \frac{a^{\nu/2}b^{1/2}}{B(1/2,\nu/2)(a+bt^2)^{1/2+\nu/2}}$$

where $B(\cdot,\cdot)$ is the Beta function. If F is distributed as $(a/b)U_1^2/U_2^2$ where $U_1 \sim \chi^2_{\nu_1}$ and $U_2 \sim \chi^2_{\nu_2}$ then F has density function

$$p(f) = \frac{a^{\nu_2}b^{\nu_1}f^{\nu_1/2-1}}{B(\nu_1,\nu_2)(a+bf)^{\nu_1/2+\nu_2/2}}$$

The null joint distribution of $\hat{\beta}$ and s^2 is

$$p(\hat{\beta}, s^2 \,|\, \beta = 0) = \int p(\hat{\beta} \,|\, \sigma^{-2}, \beta = 0) p(s^2 \,|\, \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

The integrand is

$$\begin{aligned}
& \frac{1}{(2\pi v\sigma^2)^{1/2}} \exp\left(-\frac{\hat{\beta}^2}{2v\sigma^2}\right) \\
\times \quad & \left(\frac{d}{2\sigma^2}\right)^{d/2} \frac{s^{2(d/2-1)}}{\Gamma(d/2)} \exp\left(-\frac{ds^2}{2\sigma^2}\right) \\
\times \quad & \left(\frac{d_0 s_0^2}{2}\right)^{d_0/2} \frac{\sigma^{-2(d_0/2-1)}}{\Gamma(d_0/2)} \exp\left(-\sigma^{-2}\frac{d_0 s_0^2}{2}\right) \\
= \quad & \frac{(d_0 s_0^2/2)^{d_0/2}(d/2)^{d/2} s^{2(d/2-1)}}{(2\pi v)^{1/2}\Gamma(d_0/2)\Gamma(d/2)} \\
& \sigma^{-2(1/2+d_0/2+d/2-1)} \exp\left\{-\sigma^{-2}\left(\frac{\hat{\beta}^2}{2v} + \frac{ds^2}{2} + \frac{d_0 s_0^2}{2}\right)\right\}
\end{aligned}$$

which integrates to

$$\begin{aligned}
& p(\hat{\beta}, s^2 \,|\, \beta = 0) \\
& \quad = \frac{(1/2v)^{1/2}(d_0 s_0^2/2)^{d_0/2}(d/2)^{d/2} s^{2(d/2-1)}}{D(1/2, d_0/2, d/2)} \left(\frac{\hat{\beta}^2/v + d_0 s_0^2 + ds^2}{2}\right)^{-(1+d_0+d)/2}
\end{aligned}$$

where $D()$ is the Dirichlet function.

The null joint distribution of $\tilde{t}$ and s^2 is

$$p(\tilde{t}, s^2 \,|\, \beta = 0) = \tilde{s} v^{1/2} p(\hat{\beta}, s^2 \,|\, \beta = 0)$$

which after collection of factors yields

$$\begin{aligned} p(\tilde{t}, s^2 \,|\, \beta = 0) &= \frac{(d_0 s_0^2)^{d_0/2} d^{d/2} s^{2(d/2-1)}}{B(d/2, d_0/2)(d_0 s_0^2 + d s^2)^{d_0/2+d/2}} \\ &\quad \times \frac{(d_0+d)^{-1/2}}{B(1/2, d_0/2 + d/2)} \left(1 + \frac{\tilde{t}^2}{d_0+d}\right)^{-(1+d_0+d)/2} \end{aligned}$$

This shows that $\tilde{t}$ and s^2 are independent with

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

and

$$\tilde{t} \,|\, \beta = 0 \sim t_{d_0+d}.$$

The above derivation goes through similarly with $\beta \neq 0$, the only difference being that

$$\tilde{t} \,|\, \beta \neq 0 \sim (1 + v_0/v)^{1/2} \, t_{d_0+d}.$$

The marginal distribution of $\tilde{t}$ over all the genes is therefore a mixture of a scaled t-distribution and an ordinary t-distribution with mixing proportions p and $1-p$ respectively.

5 Posterior Odds

Given the unconditional distribution of $\tilde{t}_{gj}$ and s_g^2 from the previous section, it is now easy to compute the posterior odds than any particular gene g is differentially expressed with respect to contrast β_{gj}. The odds that the gth gene has non-zero β_{gj} is

$$O_{gj} = \frac{p(\beta_{gj} \neq 0 \,|\, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0 \,|\, \tilde{t}_{gj}, s_g^2)} = \frac{p(\beta_{gj} \neq 0, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0, \tilde{t}_{gj}, s_g^2)} = \frac{p_j}{1-p_j} \frac{p(\tilde{t}_{gj} \,|\, \beta_{gj} \neq 0)}{p(\tilde{t}_{gj} \,|\, \beta_{gj} = 0)}$$

since $\tilde{t}_{gj}$ and s_g^2 are independent and the distribution of s_g^2 does not depend on β_{gj}. Substituting the density for $\tilde{t}_{gj}$ from Section 4 gives

$$O_{gj} = \frac{p_j}{1-p_j} \left(\frac{v_{gj}}{v_{gj}+v_{0j}}\right)^{1/2} \left(\frac{\tilde{t}_{gj}^2 + d_0 + d_g}{\tilde{t}_{gj}^2 \frac{v_{gj}}{v_{gj}+v_{0j}} + d_0 + d_g}\right)^{(1+d_0+d_g)/2}$$

This agrees with equation (7) of Lönnstedt and Speed (2002). In the limit for $d_0 + d_g$ large, the odds ratio reduces to

$$O_{gj} = \frac{p_j}{1-p_j} \left(\frac{v_{gj}}{v_{gj}+v_{0j}}\right)^{1/2} \exp\left(\frac{\tilde{t}_{gj}^2}{2} \frac{v_{gj}}{v_{gj}+v_{0j}}\right).$$

This expression is important for accurate computation of O_{gj} in limiting cases. Following Lönnstedt and Speed (2002), the statistic

$$B_{gj} = \log O_{gj}$$

is on a friendly scale and is useful for ranking genes in order of evidence for differential expression. Note that B_{gj} is monotonic increasing in $\tilde{t}_{gj}$ if the d_g and v_g do not vary between genes.

6 Estimation of Hyperparameters

6.1 General

The statistics B_{gj} and $\tilde{t}_{gj}$ depend on the hyperparameters in the hierachical model defined in Section 3. A fully Bayesian approach would be to allow the user to choose these parameters. This paper takes instead an empirical Bayes approach in which these parameters are estimated from the data. The purpose of this section is to develop consistent, closed form estimators for d_0, s_0 and the v_{0j} from the observed sample variances s_g^2 and moderated t-statistics $\tilde{t}_{gj}$. We estimate d_0 and s_0^2 from the s_g^2 and then estimate the v_{0j} from the $\tilde{t}_{gj}$ assuming d_0 and the p_j to be known.

Specifically, d_0 and s_0^2 are estimated by equating empirical to expected values for the first two moments of $\log s_g^2$. We use $\log s_g^2$ here instead of s_g^2 because the moments of $\log s_g^2$ are finite for any degrees of freedom and because the distribution of $\log s_g^2$ is more nearly normal so that moment estimation is likely to be more efficient. We estimate the v_{0j} by equating the order statistics of the $|\tilde{t}_{gj}|$ to their nominal values. Each order statistic of the $|\tilde{t}_{gj}|$ yields an individual estimator of v_{0j}. A final estimator of v_{0j} is obtained by averaging the estimators arising from the top $Gp_j/2$ of the order statistics where G is the number of genes.

The closed form estimators given here could be used as starting values for maximum likelihood estimation of d_0, s_0 and the v_{0j} based on the marginal distributions of the observed statistics, although that route is not followed in this paper. The large size of microarray data sets ensures that the moment estimators given here are reasonably precise without further iteration.

6.2 Estimation of d_0 and s_0

Write

$$z_g = \log s_g^2$$

Each s_g^2 follows a scaled F-distribution so z_g is distributed as a constant plus Fisher's z distribution (Johnson and Kotz, 1970, page 78). The distribution of z_g is roughly normal and has finite moments of all orders including

$$E(z_g) = \log s_0^2 + \psi(d_g/2) - \psi(d_0/2) + \log(d_0/d_g)$$

and

$$\text{var}(z_g) = \psi'(d_g/2) + \psi'(d_0/2)$$

where $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively. Write

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

Then

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

and

$$E\{(e_g - \bar{e})^2 n/(n-1) - \psi'(d_g/2)\} \approx \psi'(d_0/2)$$

We can therefore estimate d_0 by solving

$$\psi'(d_0/2) = \text{mean}\left\{(e_g - \bar{e})^2 n/(n-1) - \psi'(d_g/2)\right\} \tag{1}$$

for d_0. Although the inverse of the trigamma function is not a standard mathematical function, equation (1) can be solved very efficiently using a monotonically convergent Newton iteration as described in the appendix.

Given an estimate for d_0, s_0^2 can be estimated by

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

This estimate for s_0^2 is usually somewhat less than the mean of the s_g^2 in recognition of the skewness of the F-distribution.

Note that these estimators allow for d_g to differ between genes and therefore allow for arbitrary missing values in the expression data. Note also that any gene for which $d_g = 0$ will receive $\tilde{s}_g^2 = s_0^2$.

In the case that $\text{mean}\{(e_g - \bar{e})^2 n/(n-1) - \psi'(d_g/2)\} \leq 0$, (1) cannot be solved because the variability of the s_g^2 is less than or equal to that expected from chisquare sampling variability. In that case there is no evidence that the underlying variances σ_g^2 vary between genes so d_0 is set to positive infinity and $s_0^2 = \exp(\bar{e})$.

6.3 Estimation of v_{0j}

In this section we consider the estimation of the v_{0j} for a given j. For ease of notation, the subscript j will be omitted from $\tilde{t}_{gj}$, v_{gj}, v_{0j} and p_j for the remainder of this section.

Gene g yields a moderated t-statistic $\tilde{t}_g$. The cumulative distribution function of $\tilde{t}_g$ is

$$F(\tilde{t}_g;\, v_g, v_0, d_0 + d_g) = pF\left(\tilde{t}_g \left\{\frac{v_g}{v_g + v_0}\right\}^{1/2};\, d_0 + d_g\right) + (1-p)F\left(\tilde{t}_g;\, d_0 + d_g\right)$$

where $F(\cdot;\, k)$ is the cumulative distribution function of the t-distribution on k degrees of freedom. Let r be the rank of gene g when the $|\tilde{t}_g|$ are sorted in descending order.

We match the p-value of any particular $|\tilde{t}_g|$ to its nominal value given its rank. For any particular gene g and rank r we need to solve

$$2F(-|\tilde{t}_g|;\, v_g, v_0, d_0 + d_g) = \frac{r - 0.5}{G} \tag{2}$$

for v_0. The left-hand side is the actual p-value given the parameters and the right-hand side is the nominal value for the p-value of rank r. The interpretation is that, if a probability plot was constructed by plotting the $|\tilde{t}_g|$ against the theoretical quantiles corresponding to probabilities $(r - 0.5)/G$, (2) is the condition that a particular value of $|\tilde{t}_g|$ will lie exactly on the line of equality. The value of v_0 which solves (2) is

$$v_0 = v_g \left(\frac{\tilde{t}_g^2}{q_{\text{target}}^2} - 1 \right)$$

with

$$q_{\text{target}} = F^{-1}\left(p_{\text{target}};\, d_0 + d_g\right)$$

and

$$p_{\text{target}} = \frac{1}{p}\left\{ \frac{r - 0.5}{2G} - (1 - p)F(-|\tilde{t}_g|;\, d_0 + d_g) \right\}$$

provided that $0 < p_{\text{target}} < 1$ and $q_{\text{target}} \leq |\tilde{t}_g|$.

If $|\tilde{t}_g|$ lies above the line of equality in a t-distribution probability plot, i.e., if $F(-|\tilde{t}_g|; d_0 + d_g) < (r - 0.5)/(2G)$, then $p_{\text{target}} > 0$ and $v_0 > 0$. Restricting to those values of r for which $(r - 0.5)/(2G) < p$ ensures also that $p_{\text{target}} < 1$ so that the estimator of v_0 is defined. If $|\tilde{t}_g|$ does not lie above the line of equality then the best estimate for v_0 is zero.

To get a combined estimator of v_0, we obtain individual estimators of v_0 for $r = 1, \ldots, Gp/2$ and set v_0 to be the mean of these estimators. The combined estimator will be positive, unless none of the top $Gp/2$ values for $|\tilde{t}_g|$ exceed the corresponding order statistics of the t-distribution, in which case the estimator will be zero.

6.4 Practical Considerations: Estimation of p_j

In principle it is not difficult to estimate the proportions p_j from the data as well as the other hyperparameters. A natural estimator would be to iteratively set

$$p_j = \frac{1}{G}\sum_{i=1}^{G} \frac{O_{gj}}{1 + O_{gj}}$$

since $O_{gj}/(1 + O_{gj})$ is the estimated probability that gene g is differentially expressed, and there are other possibilities such as maximum likelihood estimation. The data however contain considerably more information about d_0 and s_0^2 than about the v_{0j} and the p_j. This is because all the genes contribute to estimation of d_0 and s_0^2 whereas only those which are differential expressed contribute to estimation of v_{0j} or p_j, and even that indirectly as the identity of the differentially expressed genes is unknown.

Any set of sample variances s_g^2 leads to useful estimates of d_0 and s_0^2. On the other hand, the estimation of v_{0j} and p_j is somewhat unstable in that estimates on the boundaries $p_j = 0$, $p_j = 1$ or $v_{0j} = 0$ have positive probability and these boundary values lead to degenerate values for the posterior odds statistics B_{gj}. Even when not on the boundary, the estimator for p_j is likely to be sensitive to the particular form of the prior distribution assumed for β_{gj} and possibly also to dependence between the genes (Ferkingstad et al, 2003).

A practical strategy to bypass these problems is to set the p_j to values chosen by the user, perhaps $p_j = 0.01$ or some other small value. Since $v_{0j}^{1/2}\sigma_g$ is the standard deviation of the log-fold-changes for differentially expressed genes, and because this standard deviation cannot be unreasonably small or large in a microarray experiment, it seems reasonable to place some prior bounds on $v_{0j}^{1/2}s_0$ as well. In the software package Limma which implements the methods in this paper, the user is allowed to place limits on the possible values for $v_{0j}^{1/2}s_0$. By default these limits are set at 0.1 and 4, chosen to include a wide range of reasonable values. For many data sets these limits do not come into play but they do prevent very small or very large estimated values for v_{0j}.

7 Inference Using Moderated t and F Statistics

As already noted, the odds ratio statistic for differential expression derived in Section 5 is monotonic increasing in $|\tilde{t}_{gj}|$ for any given j if d_g and v_g do not vary between genes. This shows that the moderated t-statistic is an appropriate statistic for assessing differential expression and is equivalent to B_{gj} as a means of ranking genes. Even when d_g and v_g do vary, it is likely that p-values from the $\tilde{t}_{gj}$ will rank the genes in very similar order to the B_{gj}.

The moderated t has the advantage over the B-statistic that B_{gj} depends on hyperparameters v_{0j} and p_j for all j as well as d_0 and s_0^2 whereas $\tilde{t}_{gj}$ depends only on d_0 and s_0^2. This means that the moderated t does not require knowledge of the proportion of differentially expressed genes, a potentially contentious parameter, nor does it make any assumptions about the magnitude of differential expression when it exists. The moderated t does require knowledge of d_0 and s_0^2, which describe the variability of the true gene-wise variances, but these hyperparameters can be estimated in stable fashion as described in Section 6.

The moderated t has the advantage over the ordinary t statistic that large statistics are less likely to arise merely from under-estimated sample variances. This is because the posterior variance $\tilde{s}_g^2$ offsets the small sample variances heavily in a relative sense while larger sample variances are moderated to a lesser relative degree. In this respect the moderated t statistic is similar in spirit to t-statistics with offset standard deviation. Provided that $d_0 < \infty$ and $d_g > 0$, the moderated t-statistic can be written

$$\tilde{t}_{gj} = \left(\frac{d_0 + d_g}{d_g}\right)^{1/2} \frac{\hat{\beta}_{gj}}{\sqrt{s_{*,g}^2 v_{gj}}}$$

where $s^2_{*,g} = s^2_g + (d_0/d_g)s^2_0$. This shows that the moderated t-statistic is proportional to a t-statistic with sample variances offset by a constant if the d_g are equal. Test statistics with offset standard deviations of the form

$$t_{*,gj} = \frac{\hat{\beta}_{gj}}{(s_g + a)\sqrt{v_{gj}}},$$

where a is a positive constant, have been used by Tusher et al (2001), Efron et al (2001) and Broberg (2003). Note that t_* offsets the standard deviation while $\tilde{t}$ offsets the variance so the two statistics are not functions of one another. Unlike the moderated t, the offset statistic $t_{*,gj}$ is not connected in any formal way with the posterior odds of differential expression and does not have an associated distributional theory.

The moderated t-statistic $\tilde{t}_{gj}$ may be used for inference about β_{gj} in a similar way to that in which the ordinary t-statistic would be used, except that the degrees of freedom are $d_g + d_0$ instead of d_g. The fact that the hyperparameters d_0 and s^2_0 are estimated from the data does not impact greatly on the individual test statistics because the hyperparameters are estimated using data from all the genes, meaning that they are estimated relatively precisely and can be taken to be known at the individual gene level.

The moderated t-statistics also lead naturally to moderated F-statistics which can be used to test hypotheses about any set of contrasts simultaneously. Appropriate quadratic forms of moderated t-statistics follow F-distributions just as do quadratic forms of ordinary t-statistics. Suppose that we wish to test all contrasts for a given gene equal to zero, i.e., $H_0 : \boldsymbol{\beta}_g = 0$. The correlation matrix of $\hat{\boldsymbol{\beta}}_g$ is $R_g = U_g^{-1} C^T V_g C U_g^{-1}$ where U_g is the diagonal matrix with unscaled standard deviations $(v_{gj})^{1/2}$ on the diagonal. Let r be the column rank of C. Let Q_g be such that $Q_g^T R_g Q_g = I_r$ and let $\mathbf{q}_g = Q_g^T \mathbf{t}_g$. Then

$$F_g = \mathbf{q}_g^T \mathbf{q}_g / r = \mathbf{t}_g^T Q_g Q_g^T \mathbf{t}_g / r \sim F_{r,d_0+d_g}$$

If we choose the columns of Q_g to be the eigenvectors spanning the range space of R_g then $Q_g^T R_g Q_g = \Lambda_g$ is a diagonal matrix and

$$F_g = \mathbf{q}_g^T \Lambda_g^{-1} \mathbf{q}_g / r \sim F_{r,d_0+d_g}.$$

The statistic F_g is simply the usual F-statistic from linear model theory for testing $\boldsymbol{\beta}_g = 0$ but with the posterior variance $\tilde{s}^2_g$ substituted for the sample variance s^2_g in the denominator.

8 Simulation Results

This section presents a simulation study comparing the moderated t-statistic with four other statistics for ranking genes in terms of evidence for differential expression. The moderated t statistic is compared with (i) ordinary fold change, equivalent to $|\hat{\beta}_{gj}|$, (ii) the ordinary t-statistic $|t_{gj}|$, (iii) Efron's idea of offsetting the standard deviations by their 90th percentile and (iv) the original Lönnstedt and Speed (2002) method as

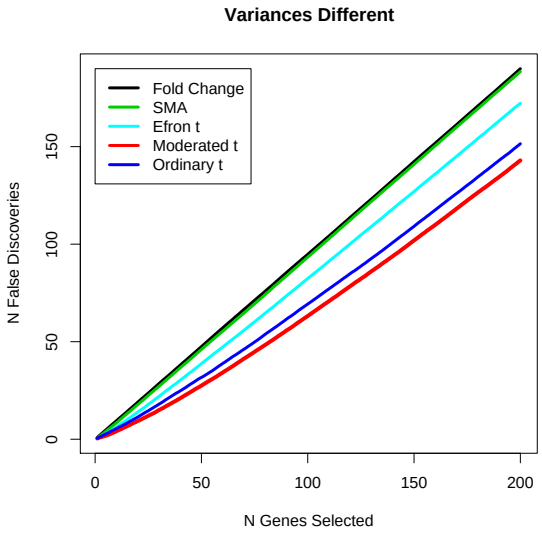

Figure 2: False discovery rates for different gene selection statistics when the true variances are very different, i.e., the residual variance dominates. The rates are means of actual false discovery rates for 100 simulated data sets.

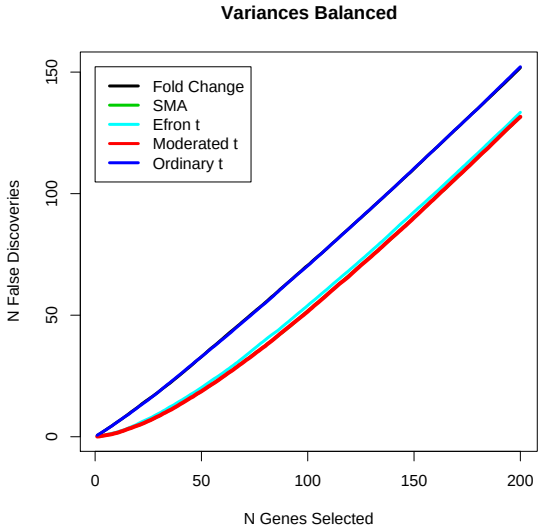

Figure 3: False discovery rates for different gene selection statistics when the true variances are somewhat different, i.e., the prior and residual degrees of freedom are balanced. The rates are means of actual false discovery rates for 100 simulated data sets.

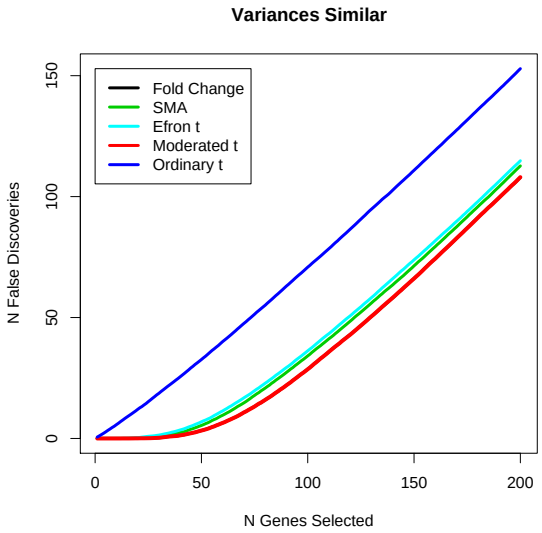

Figure 4: False discovery rates for different gene selection statistics when the true variances are somewhat different, i.e., the prior and residual degrees of freedom are balanced. The rates are means of actual false discovery rates for 100 simulated data sets.

Table 1: Area under the Receiver Operating Curve for five statistics and three simulation scenarios.

Variances	Fold Change	SMA	Efron t	Moderated t	Ordinary t
Different	0.6883	0.6888	0.7123	**0.7525**	0.7480
Balanced	0.7480	0.7592	0.7579	**0.7593**	0.7480
Similar	**0.7710**	0.7687	0.7680	**0.7710**	0.7496

Table 2: Means (standard deviations) of hyperparameter estimates for the simulated data sets. True values are $d_0/(d_0+d_g) = 0.2, 0.5, 0.9960$, $s_0^2 = 4$, $v_0 = 2$.

Variances	$d_0/(d_0+d_g)$	s_0^2	v_0, $p = 0.01$	v_0, $p = 0.02$
Different	0.2000 (0.0019)	4.0000 (0.070)	2.37 (0.21)	1.91 (0.37)
Balanced	0.5000 (0.0054)	3.9984 (0.044)	3.41 (0.38)	2.02 (0.33)
Similar	0.9901 (0.0119)	3.9922 (0.031)	3.46 (0.25)	1.98 (0.25)

implemented in the SMA package for R. For Efron's offset t-statistic we select genes in order of $|t_{E,gj}|$ where

$$t_{E,gj} = \frac{\hat{\beta}_{gj}}{(s_g + s_{0.9})\sqrt{v_{gj}}}$$

where $s_{0.9}$ is the 90th percentile of the s_g. This is corresponds to the statistic defined by Efron et al (2001). For the Lönnstedt and Speed (2002) method, genes were selected in order of the B-statistic or log-odds computed by the `stat.bay.est` function the SMA package for R available from

`http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html`.

The Lönnstedt and Speed (2002) B-statistic is equivalent here to the moderated t except for differences in the hyperparameter estimators. The moderated t-statistic was evaluated at the hyperparameter estimators derived in Section 6.

Data sets were simulated from the hierarchical model of Section 3 under three different parameter scenarios, one where the gene-wise variances are quite different, one where prior and residual degrees of freedom are balanced and one where the gene-wise variances are nearly constant. Data sets were simulated with 15000 genes, 300 or $p = 0.02$ of which were differentially expressed. All three scenarios used $d_g = 4$, $s_0^2 = 4$, $v_g = 1/3$ and $v_0 = 2$. The prior degrees of freedom was varied between the scenarios with $d_0 = 1$ for the scenario with very different variances, $d_0 = 4$ to balance the prior and residual degrees of freedom and $d_0 = 1000$ to make the true variances almost identical. A hundred data sets were simulated for each scenario.

Figures 2-4 plot the average false discovery rates for the five statistics and Table 1 gives the areas under the receiver operating curves (ROCs). The moderated t-statistic has the lowest false discovery rate and the highest area under the ROC for all three scenarios. The straight fold change does equally well, as expected, when the true variances are nearly constant. The SMA B-statistic and Efron's t statistic do nearly as well as the moderated t statistic when the prior and residual degrees of freedom are balanced but they are less able to adapt to other arrangements for the true variances. Although the moderated t can be expected to perform better than the fold change and the ordinary t-statistic on data simulated under the assumed hierarchical model, the simulations show that this advantage is realized even though the hyperparameters need to be estimated.

Table 2 give means and standard deviations of the hyperparameter estimates for the simulated data sets. The estimates for $d_0/(d_0 + d_g)$ and s_0^2 were very accurate. The estimator for v_0 was nearly unbiased when p was set to the true proportion of differentially expressed genes ($p = 0.02$) but was somewhat over-estimated when the proportion was set to a lower value ($p = 0.01$). As expected the estimator for v_0 is somewhat more variable than that of $d_0/(d_0 + d_g)$ or s_0^2. The results shown in Table 2 are not affected by the prior limits on $v_0 s_0^2$ discussed in Section 6.

9 Data Examples

9.1 Swirl

Consider the Swirl data set which is distributed as part of the marrayInput package for R (Dudoit and Yang, 2003). The experiment was carried out using zebrafish as a model organism to study the early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. The main goal of the Swirl experiment is to identify genes with altered expression in the Swirl mutant compared to wild-type zebrafish. The experiment used four arrays in two dye-swap pairs. The microarrays used in this experiment were printed with 8448 probes (spots) including 768 control spots. The hybridized microarrays were scanned with an Axon scanner and SPOT image analysis software (Buckley, 2000) was used to capture red and green intensities for each spot. The data was normalized using print-tip loess normalization and between arrays scale normalization using the LIMMA package (Smyth, 2003). Loess normalization used window span 0.3 and three robustifying iterations.

For this data all genes have $d_g = 3$ and $v_g = 1/4$. The estimated prior degrees of freedom are $d_0 = 4.17$ showing that posterior variances will be balanced between the prior and sample variances with only slightly more weight to the former. The estimated prior variance is $s_0^2 = 0.0509$ which is less then the mean variance at 0.109 but more than the median at 0.047. The estimated unscaled variance for the contrast is $v_0 = 22.7$, meaning that the standard deviation of the log-ratio for a typical gene is $(0.0509)^{1/2}(22.7)^{1/2} = 1.07$, i.e., genes which are differentially expressed typically change by about two-fold. The prior limits on $v_0 s_0^2$ discussed in Section 6 do not come into play for this data. The proportion of differentially expressed genes was set to $p = 0.01$ following the default in the earlier SMA software. This value seems broadly realistic for single-gene knock-out vs wild-type experiments but is probably conservative for other experiments where more differential expression is expected.

Table 3 shows the top 30 genes as ranked by the B-statistic. The table includes the fold change (or M-value), the ordinary t-statistic, the moderated t-statistic and the B-statistic for the each gene. The moderated t and the B-statistics are evaluated at the hyperparameter estimators derived in Section 6. There are no missing values for this data so the two statistics necessarily give the same ranking of the genes. The ordinary t-statistics are on 3 degrees of freedom while the moderated t are on 7.17. The moderated t-statistic method here ranks both copies of the knock-out gene BMP2 first

Table 3: Top 30 genes from the Swirl data

ID	Name	M-value	Ord t	Mod t	B
control	BMP2	-2.21	-23.94	-21.1	7.96
control	BMP2	-2.30	-20.20	-20.3	7.78
control	Dlx3	-2.18	-21.03	-20.0	7.71
control	Dlx3	-2.18	-20.09	-19.6	7.62
fb94h06	20-L12	1.27	30.23	14.1	5.78
fb40h07	7-D14	1.35	17.39	13.5	5.54
fc22a09	27-E17	1.27	21.11	13.4	5.48
fb85f09	18-G18	1.28	20.23	13.4	5.48
fc10h09	24-H18	1.20	28.30	13.2	5.40
fb85a01	18-E1	-1.29	-17.39	-13.1	5.32
fb85d05	18-F10	-2.69	-9.23	-13.0	5.29
fb87d12	18-N24	1.27	16.76	12.8	5.22
control	Vox	-1.26	-17.22	-12.8	5.20
fb85e07	18-G13	1.23	18.26	12.8	5.18
fb37b09	6-E18	1.31	14.02	12.4	5.02
fb26b10	3-I20	1.09	39.13	12.4	4.97
fb24g06	3-D11	1.33	13.26	12.3	4.96
fc18d12	26-F24	-1.25	-14.42	-12.2	4.89
fb37e11	6-G21	1.23	14.48	12.0	4.80
control	fli-1	-1.32	-12.31	-11.9	4.76
control	Vox	-1.25	-13.24	-11.9	4.71
fb32f06	5-C12	-1.10	-18.52	-11.7	4.63
fb50g12	9-L23	1.16	15.08	11.7	4.63
control	vent	-1.40	-10.90	-11.7	4.62
fb23d08	2-N16	1.16	14.95	11.6	4.58
fb36g12	6-D23	1.12	13.63	11.0	4.27
control	vent	-1.41	-9.34	-10.8	4.13
control	vent	-1.37	-8.98	-10.5	3.91
fb22a12	2-I23	1.05	11.96	10.2	3.76
fb38a01	6-I1	-1.82	-7.54	-10.2	3.75

and both copies of Dlx3, which is a known target of BMP2, second. Neither the fold change nor the ordinary t-statistic do this. In general the moderated methods give a more predictable ranking of the control genes.

9.2 ApoAI

These data are from a study of lipid metabolism by Callow et al (2000) and are available from

`http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html`.

The apolipoprotein AI (ApoAI) gene is known to play a pivotal role in high density lipoprotein (HDL) metabolism. Mouse which have the ApoAI gene knocked out have very low HDL cholesterol levels. The purpose of this experiment is to determine how ApoAI deficiency affects the action of other genes in the liver, with the idea that this will help determine the molecular pathways through which ApoAI operates.

The experiment compared 8 ApoAI knockout mice with 8 normal C57BL/6 mice, the control mice. For each of these 16 mice, target mRNA was obtained from liver tissue and labelled using a Cy5 dye. The RNA from each mouse was hybridized to a separate microarray. Common reference RNA was labelled with Cy3 dye and used for all the arrays. The reference RNA was obtained by pooling RNA extracted from the 8 control mice. Although still a simple design, this experiment takes us outside the replicated array structure considered by Lönnstedt and Speed (2002).

Intensity data was captured from the array images using SPOT software and the data was normalized using print-tip loess normalization. The normalized intensities were analyzed for differential expression by fitting a two-parameter linear model, the parameter of interest measuring the difference between the ApoAI knockout line and the control mice.

The residual degrees of freedom are $d_g = 14$ for most genes but some have d_g as low as 10. The estimated prior degrees of freedom are $d_0 = 3.7$. The residual degrees of freedom are relatively large here so the sample variances will be shrunk only slightly and the moderated and ordinary t-statistics will differ substantially only when the sample variance is unusually small. The prior variance is $s_0^2 = 0.048$ which is somewhat less than the median sample variance at 0.064. The unscaled variance for the contrasts of interest is estimated to be $v_0 = 3.4$ meaning that the typical fold change for differentially expressed genes is estimated to be about 1.3. Prior limits on $v_0 s_0^2$ do not come into play for these data.

Table 4 shows the top 15 genes as ranked by the B-statistic for the parameter of interest. The moderated t rank the genes in the same order even though there are a few missing values for this data. The top gene is ApoAI itself which is heavily down-regulated as expected. Several of the other genes are closely related to ApoAI. The top eight genes here have been confirmed to be differentially expressed in the knockout versus the control line (Callow et al, 2000). For these data the top eight genes stand out clearly from the other genes and all methods clearly separate these genes from the

Table 4: Top 15 genes from the ApoAI data

Annotation	M-value	Ord t	Mod t	B
Apo AI, lipid-Img	-3.17	-23.1	-24.0	15.96
EST highly similar to Apolipoprotein A-I precursor, lipid-UG	-3.05	-11.8	-12.9	11.35
Catechol O-Methyltransferase, membrane-bound, brain-Img	-1.85	-11.8	-12.5	10.98
EST similar to C-5 Sterol Desaturase, lipid-UG	-1.03	-13.0	-11.9	10.52
Apo CIII, lipid-Img	-0.93	-10.4	-9.9	8.65
EST highly similar to Apolipoprotein C-III precursor, lipid-UG	-1.01	-9.0	-9.1	7.67
EST	-0.98	-9.1	-9.1	7.66
Similar to yeast sterol desaturase, lipid-Img	-0.95	-7.2	-7.5	5.55
EST similar to fatty acid-binding protein, epidermal, lipid-UG	-0.57	-4.4	-4.6	0.63
Clone ID 317638	-0.37	-4.3	-4.0	-0.47
APXL2, 5q-Img	-0.42	-4.0	-4.0	-0.56
Estrogen rec	0.42	4.0	3.9	-0.60
Caspase 7, heart-Img	-0.30	-4.6	-3.9	-0.64
Psoriasis-associated fatty acid binding protein, lipid-Img	-0.84	-3.6	-3.9	-0.69
Fatty acid-binding protein, epidermal, lipid-UG	-0.64	-3.6	-3.8	-0.84

others. The B-statistic for example drops from 5.55 to 0.63 from the 8th to the 9th ranked gene.

10 Software

The methods described in this paper, including linear models and contrasts as well as moderated t and F statistics and posterior odds, are implemented in the software package Limma for the R computing environment (Smyth et al, 2003). Limma is part of the Bioconductor project at http://www.bioconductor.org (Gentleman et al, 2003). The Limma software has been tested on a wide range of microarray data sets from many different facilities and has been used routinely at the author's institution since the middle of 2002.

Colophon

The author thanks Terry Speed, Ingrid Lönnstedt and Yu Chuan Tai for valuable discussions and for comments on an earlier version of this manuscript. Thanks also to Suzanne Thomas for assistance with the figures. This research was supported by NHMRC Grants 257501 and 257529.

Appendix: Inversion of Trigamma Function

In this appendix we solve $\psi'(y) = x$ for y where $x > 0$ by deriving a Newton iteration with guaranteed and rapid convergence.

Define $f(y) = 1/\psi'(y)$. The function f is nearly linear and convex for $y > 0$, satisfying $f(0) = 0$ and asymptoting to $f(y) = y - 0.5$ as $y \to \infty$. The first derivative

$$f'(y) = -\frac{\psi''(y)}{\psi'(y)^2}$$

is strictly increasing from $f'(0) = 0$ to $f'(\infty) = 1$. This means that the Newton iteration to solve $f(y) = z$ for y is monotonically convergent provided that the starting value y_0 satisfies $f(y_0) \geq z$. Such a starting value is provided by $y_0 = 0.5 + z$.

The complete Newton iteration to solve $\psi'(y) = x$ is as follows. Set $y_0 = 0.5 + 1/x$. Then iterate $y_{i+1} = y_i + \delta_i$ with $\delta_i = \psi'(y_i)\{1 - \psi'(y_i)/x\}/\psi''(y_i)$ until $-\delta_i/y < \epsilon$ where ϵ is a small positive number. The step δ_i is strictly negative unless $\psi'(y_i) = x$. Using 64-bit double precision arithmetic, $\epsilon = 10^{-8}$ is adequate to achieve close to machine precision accuracy. To avoid overflow or underflow in floating point arithmetic, we can set $y = 1/\sqrt{x}$ when $x > 10^7$ and $y = 1/x$ when $x < 10^{-6}$ instead of performing the iteration. Again these choices are adequate for nearly full precision in 64-bit arithmetic.

References

Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biology* **4**: R41.

Buckley, M. J. (2000). Spot User's Guide. CSIRO Mathematical and Information Sciences, Sydney, Australia. http://www.cmis.csiro.au/iap/Spot/spotmanual.htm.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* **10**, 2022–2029.

Chu, T.-M., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* **176**, 35–51.

Cui, X., and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.1–210.9.

Dudoit, S., and Yang, Y. H. (2003). Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S. Garrett, R. A.

Irizarry and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York. pp. 73-101.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Annals of Statistics* **31**, 366–378.

Efron B., Tibshirani, R., Storey J. D., and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Ferkingstad, E., Langaas, M., and Lindqvist, B. (2003). Estimating the proportion of true null hypotheses, with application to DNA microarray data. Preprint Statistics No. 4/2003, Norwegian University of Science and Technology, Trondheim, Norway. http://www.math.ntnu.no/preprint/

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis, with discussion. *TEST* **12**, 1–78.

Gentleman, R., Bates, D., Bolstad, B., Carey, V., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., C., Maechler, M., Rossini, A. J., Sawitzki, G., Smyth, G. K., Tierney, L., Yang, J. Y. H., and Zhang, J. (2003). Bioconductor: a software development project. Technical Report November 2003, Department of Biostatistics, Harvard School of Public Health, Boston.

Huber, W., von Heydebreck, A., Sltmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Society* **97**, 88-99.

Irizarry, R. A., Bolstad, B. M., Francois Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003), Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**(4):e15.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* **29**, 389–395.

Johnson, N. L., and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions – 2*. Wiley, New York.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*. To appear.

Kerr, M. K., and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98**, 31–36.

Lönnstedt, I. (2001). *Replicated Microarray Data*. Licentiate Thesis, Department of Mathematics, Uppsala University.

Lönnstedt, I., Grant, S., Begley, G., and Speed, T. P. (2003). Microarray analysis of two interacting treatments: a linear model and trends in expression over time. To appear.

Lönnstedt, I., and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Nature Genetics Editors (eds.) (2003). Chipping Forecast II. *Nature Genetics Supplement* **32**, 461–552.

Newton, M.A. and Kendziorski, C. M. (2003). Parametric empirical Bayes methods for microarrays. In: *The analysis of gene expression data: methods and software*. Eds. G. Parmigiani, E. S. Garrett, R. Irizarry and S. L. Zeger, Springer Verlag, New York. To appear.

Newton, M. A., Kenziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (eds.) (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York.

Smyth, G. K., Thorne, N. P., and Wettenhall, J. (2003). *Limma: Linear Models for Microarray Data User's Guide*. Software manual available from http://www.bioconductor.org.

Smyth, G. K., and Speed, T. P. (2003). Normalization of cDNA microarray data. In: *METHODS: Selecting Candidate Genes from DNA Array Screens: Application to Neuroscience*, D. Carter (ed.). Methods Volume 31, Issue 4, December 2003, pages 265–273.

Smyth, G. K., Yang, Y.-H., Speed, T. P. (2003). Statistical issues in microarray data analysis. In: *Functional Genomics: Methods and Protocols*, M. J. Brownstein and A. B. Khodursky (eds.), Methods in Molecular Biology Volume 224, Humana Press, Totowa, NJ, pages 111–136.

Speed, T. P. (ed.) (2003). *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Boca Raton.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* **98**, 5116–5121.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

Yang, Y. H., and Speed, T. P. (2003). Design and analysis of comparative microarray experiments. In T. P. Speed (ed.), *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC Press, pages 35–91.