Statistical Applications in Genetics and Molecular Biology

Manuscript 1762

A Mixture-Model Approach for Parallel Testing for Unequal Variances

Haim Y. Bar, *Cornell University*
James G. Booth, *Cornell University*
Martin T. Wells, *Cornell University*

A Mixture-Model Approach for Parallel Testing for Unequal Variances

Haim Y. Bar, James G. Booth, and Martin T. Wells

Abstract

Testing for unequal variances is usually performed in order to check the validity of the assumptions that underlie standard tests for differences between means (the t-test and anova). However, existing methods for testing for unequal variances (Levene's test and Bartlett's test) are notoriously non-robust to normality assumptions, especially for small sample sizes. Moreover, although these methods were designed to deal with one hypothesis at a time, modern applications (such as to microarrays and fMRI experiments) often involve parallel testing over a large number of levels (genes or voxels). Moreover, in these settings a shift in variance may be biologically relevant, perhaps even more so than a change in the mean. This paper proposes a parsimonious model for parallel testing of the equal variance hypothesis. It is designed to work well when the number of tests is large; typically much larger than the sample sizes. The tests are implemented using an empirical Bayes estimation procedure which `borrows information' across levels. The method is shown to be quite robust to deviations from normality, and to substantially increase the power to detect differences in variance over the more traditional approaches even when the normality assumption is valid.

KEYWORDS: empirical Bayes, EM algorithm, shrinkage estimation

Author Notes: Prof. Booth's research was partially supported by an NSF grant, NSF-DMS 0805865. Prof. Wells' research was partially supported by NIH grants R01-GM083606 and P60-MD08005.

1 Introduction and Motivation

Research questions are often framed in terms of the effect a treatment has on a response. When comparing two conditions (say control and treatment) the question is typically interpreted in terms of the difference between the two means. When the response is continuous, the most widely-used test to detect the effect of the treatment is the two-sample t-test. In this context, a test for unequal variances can be performed to assess the validity of the equal variance assumption. Unlike the t-test, tests for equality of variances are notoriously non-robust, as highlighted by George Box's famous quote, "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" (Box, 1953). Reviews of the literature on testing equality of variances can be found in Boos and Brownie (1989), Boos and Brownie (2004) and Gastwirth et al. (2010).

Increasingly, however, there are new insights that suggest that biological variance plays an important role in determining cellular and organismal processes. This article is concerned with problems in which testing for unequal variances is of scientific importance in its own right. Furthermore, the focus is on situations in which a large number of parallel tests are conducted. A parsimonious model and empirical Bayes estimation procedure is developed that 'borrows strength' across the levels being tested. This results not only in increased power over standard tests when the normality assumption holds, but also in substantially improved performance when it does not. The wide applicability of the approach is illustrated using four different types of data sets: gene expression, gene methylation, functional Magnetic Resonance Imaging (fMRI), and metabolomics data. In these settings changes in the variance under the treatment are often biologically relevant. Moreover, failing to account for unequal variances may undermine the performance of methods for detecting for changes in the mean.

As a specific example, consider a simple Pavlovian-type learning experiment in which the response is measured in terms of volume of blood flowing through voxels in the brain. Both control and treatment groups receive a simple visual signal in regular intervals. The subjects in the treatment group also receive an auditory stimulus in addition to the visual signal. Since the stimulus does not require complicated cognitive processing, it is conceivable that the overall mean response levels will not differ between the groups. However, in the treatment group, in preparation for the audio signal relevant areas in the brain might have smaller variability, to ensure the availability of the necessary level of blood for the anticipated task. Similarly, areas not involved in processing the audio signal might exhibit increased variability among the treatment subjects.

Variation, in genetic and phenotypic terms, has been thought to be a component of population fitness and adaptability. One way to interpret the association between expression variance and phenotype is to consider changes in pathways. If the genes in a particular pathway have very low variance, a natural interpretation is that those genes are highly constrained. Ho et al. (2008) report that they "found that changes in expression variability are associated with changes in co-expression patterns. Therefore, differential variability is potentially an important manifestation of changes in gene regulation." Hansen et al. (2011) say that "the increased across-sample variability in methylation within the cancer samples of each tumor type compared to normal was even more striking than the differences in mean methylation." Other recent examples where biological sources of variation play an important role in determining cellular and organismal processes can be found in Levsky et al. (2002), Ozbudak et al. (2002), Ravasi et al. (2002), Colman-Lerner et al. (2005), Cai et al. (2006), Mar et al. (2006), Manolio et al. (2009), Eichler et al. (2010), Feinberg and Irizarry (2010), Mar et al. (2011), and Marko et al. (2011). The recent methodology proposed by Mar et al. (2011) for assessing the variance of gene expression uses a one at a time analysis of the coefficient of variation. Mar et al. (2011) compute the coefficient of variation for each gene by dividing the standard deviation of its expression measures across a sample population by its average expression. They then designate low variance genes as those falling below the lower 25th percentile of the genome-wide coefficient of variation distribution based on all donors and high variance genes as those above the 75th percentile; those genes in the range between the 25th and 75th percentile they refer to as the mid variability gene set.

In the microarray context it is now widely recognized that methods that borrow strength across genes, by assuming that the gene-specific variances come from a common distribution, are more powerful for detecting mean treatment effects (Smyth, 2004; Bar et al., 2010; Hwang and Liu, 2010). However, these methods all assume variance homogeneity across conditions. This article develops a new model-based approach to parallel testing for unequal variances that complements the existing methods for detecting changes in the mean. Although the methodology can be applied in a variety of settings, for simplicity of exposition we use terminology from the microarray literature, and so the parallel tests concern variability in gene-specific expression in arrays based on samples from control and treatment groups.

Our model assumes that the ratio of the sample variances from the control and treatment groups arises from a three components mixture: a null component in which the ratio is proportional to an F-statistic; and two non-null groups representing inflated and deflated variance in the treatment group relative to the control. The three component mixture is identified by a latent multinomial random variable

which is treated as missing data when fitting the model via the EM algorithm. Two variants of the model are considered: one in which the inflation/deflation factors are constant across all the parallel tests; and one in which they are assumed to come from a lognormal distribution. Genes are declared as non-null if their posterior null probability is less than a predefined threshold. Alternatively, frequentist inference can be conducted by controlling the false discovery rate using the estimated null distribution.

Our approach to determining high and low variance is in line with a growing literature on empirical Bayesian analysis of high dimensional data (see Efron, 2008 and Bar et al., 2010). The hierarchical nature of our method yields shrinkage estimation which results in high power and accuracy, while maintaining a low false discovery rate. Furthermore, we show in Section 5 that the inference based on our approach is quite robust to deviations from normality assumptions.

The paper is organized as follows. The mixture model is defined in Section 2. Section 3 outlines the details of the EM algorithm. The empirical Bayes and frequentist inference procedures are described in Section 4. A simulation study demonstrating the improved power, robustness, and accuracy of the method relative to 'one gene at a time' approach is discussed in Section 5. Section 6 presents results from four case studies and some concluding remarks are given in Section 7.

2 The Mixture Model

Denote the (normalized) response for gene g in array j under condition i by y_{ijg}, and suppose that, given the gene-specific variances, σ^2_{1g} and σ^2_{2g},

$$y_{ijg} \sim N(\mu_{ig}, \sigma^2_{ig}) \tag{1}$$

independently, for all i, j and g, where $i = 1$ for arrays in the control group and $i = 2$ for the treatment group, $j = 1, \ldots, n_{ig}$, and $g = 1, \ldots, G$. Typically G is in the hundreds or thousands, whereas the sample sizes, n_{ig}, are much smaller, often only in the single digits.

The sample variance for gene g in condition i is given by

$$s^2_{ig} = \sum_{j=1}^{n_{ig}} (y_{ijg} - \bar{y}_{i\cdot g})^2 / d_{ig}\,, \tag{2}$$

where $d_{ig} = n_{ig} - 1$. It follows from the normality assumption (1) that the ratio of variances in the control and treatment samples is proportional to a central F-statistic;

that is,

$$r_g|\rho_g \sim \rho_g \frac{\chi^2_{d_{2g}}/d_{2g}}{\chi^2_{d_{1g}}/d_{1g}}, \tag{3}$$

where $r_g = s^2_{2g}/s^2_{1g}$ and $\rho_g = \sigma^2_{2g}/\sigma^2_{1g}$. In order to classify the genes as having the same, inflated or deflated variance under treatment we suppose that each ratio, ρ_g, $g = 1,\ldots,G$, is drawn from a three components mixture with probability vector, $\boldsymbol{p} = (p_0, p_1, p_2)$. Associated with each gene is a trivariate latent indicator vector $\boldsymbol{\delta}_g = (\delta_{0g}, \delta_{1g}, \delta_{2g})$ distributed as multinomial$(1, \boldsymbol{p})$ which determines whether the variance in the treatment group is null, inflated or deflated with respect to the control group. More specifically,

$$\rho_g|\delta_g, \lambda_g \sim \tau\lambda_g^{\delta_{1g}-\delta_{2g}}, \tag{4}$$

where $\lambda_g > 0$ is a gene-specific inflation/deflation factor, and the parameter τ allows for the incorporation of fixed covariate effects into the model. In the simplest case, with no covariates, τ represents a constant multiplicative difference between the variances in the control and treatment groups which is often noticeable in real data. For example, in fMRI data the stimulus presented to the treatment group may affect subjects' overall brain activity, and not just regions in the brain that are associated with the task.

We consider two variants of the model: a *fixed inflation factor* model in which $\lambda_g \equiv \lambda$, where λ is constant across all genes; and a *random inflation factor* model in which the λ_g's are assumed to come from a lognormal distribution. These assumptions both lead to inferences about the variance ratios that borrow strength across the genes, resulting in greater power to detect inflated or deflated variance under treatment.

The assumption of a lognormal distribution for λ_g can be motivated from the perspective of classical shrinkage estimation (James and Stein, 1961) and its connection to BLUPs arising in linear mixed models (Efron and Morris, 1975). Specifically, consider the variable $x_g \equiv \log(r_g)$. Equations (3) and (4) imply that

$$x_g = \log\tau + (\delta_{1g} - \delta_{2g})\log\lambda_g + \xi_{2g} - \xi_{1g} \tag{5}$$

where $\xi_{ig} = \log(\chi^2_{d_{ig}}/d_{ig})$, $i = 1,2$, have known mean and variance given by $E(\xi_{ig}) = \psi(d_{ig}/2) - \log(d_{ig}/2)$ and $\mathrm{Var}(\xi_{ig}) = \psi'(d_{ig}/2)$, ψ and ψ' being the digamma and trigamma functions, respectively. Using independence and applying the delta

method implies that $\xi_{2g} - \xi_{1g}$ is approximately normal with mean and variance given by

$$\theta_g = \psi(d_{2g}/2) - \log(d_{2g}/2) - \psi(d_{1g}/2) + \log(d_{1g}/2)$$

and

$$\kappa_g^2 = \psi'(d_{1g}/2) + \psi'(d_{2g}/2)\,.$$

Thus, if $\log\lambda_g \sim N(\theta, \kappa^2)$, equation (5) has the form of a mixture of linear mixed models, and shrinkage estimates of individual components of $\log\lambda_g$ can be estimated by their posterior expectations given the observed x_g, $g = 1, \ldots, G$ (Efron and Morris, 1975).

3 The EM Algorithm

3.1 Complete data log-likelihood

Regarding the latent indicator vector $\boldsymbol{\delta}_g$ as missing data, we obtain the complete data log-likelihood to implement the EM algorithm.

For the fixed inflation factor model where $\lambda_g \equiv \lambda$, the complete data log likelihood (omitting terms that do not depend on unknown parameters) is obtained directly from the identities (3) and (4) as

$$\begin{aligned}
\sum_{g=1}^{G} \ell_F(r_g) &= \sum_{g=1}^{G} \left\{ \sum_{k=0}^{2} \delta_{kg} \log p_k + \frac{d_{1g}}{2} \log\left(\tau\lambda^{\,\delta_{1g}-\delta_{2g}}\right) \right. \\
&\qquad \left. - \frac{d_{1g}+d_{2g}}{2} \log\left(\tau\lambda^{\,\delta_{1g}-\delta_{2g}} + r_g d_{2g}/d_{1g}\right) \right\} \\
&= \sum_{g=1}^{G}\sum_{k=0}^{2} \delta_{kg} \log p_k + \sum_{g=1}^{G} \frac{d_{1g}}{2} \left\{ \log\tau + (\delta_{1g} - \delta_{2g}) \log\lambda \right\} \\
&\qquad - \sum_{g=1}^{G} \frac{d_{1g}+d_{2g}}{2} \left\{ \delta_{0g} \log\left(\tau + r_g d_{2g}/d_{1g}\right) \right. \\
&\qquad\qquad + \delta_{1g} \log\left(\tau\lambda + r_g d_{2g}/d_{1g}\right) \\
&\qquad\qquad \left. + \delta_{2g} \log\left(\tau/\lambda + r_g d_{2g}/d_{1g}\right) \right\}\,. \qquad (6)
\end{aligned}$$

For the random inflation factor model, using the normal approximation to the log chi-squared distribution, and the mixed linear model representation in (5), we obtain

$$\begin{aligned}\sum_{g=1}^{G}\ell_R(x_g) &= \sum_{g=1}^{G}\sum_{k=0}^{2}\delta_{kg}\log p_k - \frac{1}{2}\sum_{g=1}^{G}\log[(\delta_{1g}-\delta_{2g})^2\kappa^2+\kappa_g^2)] \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\frac{[x_g-\mu_g-(\delta_{1g}-\delta_{2g})\theta]^2}{(\delta_{1g}-\delta_{2g})^2\kappa^2+\kappa_g^2} \\ &= \sum_{g=1}^{G}\sum_{k=0}^{2}\delta_{kg}\log p_k \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\left[\delta_{0g}\log(\kappa_g^2)+\delta_{1g}\log(\kappa^2+\kappa_g^2)+\delta_{2g}\log(\kappa^2+\kappa_g^2)\right] \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\left[\delta_{0g}\frac{[x_g-\mu_g]^2}{\kappa_g^2}+\delta_{1g}\frac{[x_g-\mu_g-\theta]^2}{\kappa^2+\kappa_g^2}+\delta_{2g}\frac{[x_g-\mu_g+\theta]^2}{\kappa^2+\kappa_g^2}\right],\end{aligned} \tag{7}$$

where $\mu_g = \log\tau + \theta_g$ is the expected value of x_g in the null case (i.e. when $\delta_{0g} = 1$).

3.2 The E-step

The E-step of the EM algorithm involves taking the expectation of the complete data log-likelihood conditional on the observed data. In the context of our mixture model, strict implementation of the E-step requires evaluating the expectation of all components of the complete data log-likelihood that are functions of the latent indicator, $\boldsymbol{\delta}_g$, $g = 1,\ldots,G$. In particular, if the complete data likelihood is linear in the latent indicator, as in (6) the E-step reduces to evaluating the posterior probabilities,

$$pr(\delta_{kg}=1|r_g) = \frac{p_k L_k(r_g)}{\sum_{l=0}^{2} p_l L_l(r_g)}, \tag{8}$$

at the current (iteration) parameter estimates, where $L_k(r_g) = \exp\{\ell_F(r_g)\}$ with $\delta_{kg} = 1$. The same argument holds for the random inflation factor model with ℓ_R replacing ℓ_F in the posterior probability formula (8).

3.3 The M-Step

Let φ denote the complete vector of model parameters, and let

$$Q(\varphi,\varphi^{(t)}) = E_{\varphi^{(t)}}[\ell(\{r_g\})]$$

denote the Q-function obtained by substituting the estimated posterior probabilities, $\hat{\delta}_{kg}^{(t)}$, after iteration t in (6) or (7). The M-step at the $(t+1)$st iteration involves maximization of $Q(\varphi, \varphi^{(t)})$ with respect to each parameter in φ. That is,

$$\varphi^{(t+1)} = \arg\max_{\varphi} Q(\varphi, \varphi^{(t)}) .$$

Maximization of the Q-function with respect to the multinomial probabilities is the same for both fixed and random inflation factor models, the update at iteration $t+1$ being

$$\hat{p}_k^{(t+1)} = \frac{1}{G} \sum_{g=1}^{G} \hat{\delta}_{kg}^{(t)} . \tag{9}$$

The other parameters updates depend on the assumptions regarding the inflation factors.

3.3.1 M-Step: Fixed Inflation Factor

Differentiating (6) results in the following update equations for τ and λ, respectively:

$$\sum_{g=1}^{G} \frac{d_{1g} d_{2g} \left(r_g - \tau \lambda^{\hat{\delta}_{1g} - \hat{\delta}_{2g}} \right)}{d_{2g} r_g + d_{1g} \tau \lambda^{\hat{\delta}_{1g} - \hat{\delta}_{2g}}} = 0, \tag{10}$$

and

$$\sum_{g=1}^{G} \frac{d_{1g} d_{2g} (\hat{\delta}_{1g} - \hat{\delta}_{2g}) \left(r_g - \tau \lambda^{\hat{\delta}_{1g} - \hat{\delta}_{2g}} \right)}{d_{2g} r_g + d_{1g} \tau \lambda^{\hat{\delta}_{1g} - \hat{\delta}_{2g}}} = 0 . \tag{11}$$

If $\hat{\delta}_{1g} = \hat{\delta}_{2g} = 0$ for all g, set $\hat{\lambda} = 1$.

3.3.2 M-Step: Random Inflation Factor

Differentiating (7) results in the update equations for τ, θ and κ^2:

$$\log \hat{\tau} = \frac{\sum_{g=1}^{G} \left[\frac{\delta_{0g}(x_g - \theta_g)}{\kappa_g^2} + \frac{(\delta_{1g} + \delta_{2g})(x_g - \theta_g) + (\delta_{2g} - \delta_{1g})\theta}{\kappa^2 + \kappa_g^2} \right]}{\sum_{g=1}^{G} \left(\frac{\delta_{0g}}{\kappa_g^2} + \frac{\delta_{1g} + \delta_{2g}}{\kappa^2 + \kappa_g^2} \right)} , \tag{12}$$

$$\hat{\theta} = \frac{\sum_{g=1}^{G}(\delta_{1g} - \delta_{2g})\frac{x_g - \mu_g}{\kappa^2 + \kappa_g^2}}{\sum_{g=1}^{G}\frac{\delta_{1g} + \delta_{2g}}{\kappa^2 + \kappa_g^2}}, \tag{13}$$

and

$$\hat{\kappa}^2 = \frac{\sum_{g=1}^{G}\delta_{1g}\left[(x_g - \mu_g - \theta)^2 - \kappa_g^2\right] + \delta_{2g}\left[(x_g - \mu_g + \theta)^2 - \kappa_g^2\right]}{\sum_{g=1}^{G}(\delta_{1g} + \delta_{2g})}, \tag{14}$$

and $\hat{\theta} = \hat{\kappa} = 0$ if $\hat{\delta}_{1g} = \hat{\delta}_{2g} = 0$ for all g.

4 Inference

4.1 The Frequentist and Empirical Bayes Procedures

Our model based approach allows us to assess the null status of a gene, either using a frequentist procedure based on false discovery rate (FDR, Benjamini and Hochberg, 1995); or using empirical Bayes inference via the posterior null probabilities.

Under the fixed inflation factor model the statistic, r_g/τ has an F-distribution under the null. In this case the frequentist p-value for gene g is equal to $pr(\tau F < r_{g,obs})$ if $r_g/\tau < 1$ and $pr(\tau F > r_{g,obs})$ if $r_g/\tau > 1$, where $F \sim F(d_{2g}, d_{1g})$. For the random inflation factor model the corresponding p-value is given by $pr\{|Z| > (x_{g,obs} - \mu_g)/\kappa_g\}$, where Z is a standard normal variate.

The empirical Bayes approach is to classify genes based on the estimated posterior probabilities, $\hat{\delta}_{kg}$, $k = 0, 1, 2$. Thus, a gene is declared non-null if either $\hat{\delta}_{1g}$ or $\hat{\delta}_{2g}$ exceed a given threshold.

4.2 Shrinkage Estimation

For the random factor model, the posterior probability of $\delta_{1g} = 1$ can be rewritten in the form

$$pr(\delta_{1g} = 1|x_g) = \frac{1}{\frac{p_0 \cdot L_0(x_g)}{p_1 \cdot L_1(x_g)} + 1 + \frac{p_2 \cdot L_2(x_g)}{p_1 \cdot L_1(x_g)}},$$

where the ratio, L_0/L_1, is given by

$$\begin{aligned}\frac{L_0(x_g)}{L_1(x_g)} &= \frac{(2\pi\kappa_g^2)^{-1/2}\exp\left\{-(x_g-\mu_g)^2/2\kappa_g^2\right\}}{[2\pi(\kappa^2+\kappa_g^2)]^{-1/2}\exp\left\{-(x_g-\mu_g-\theta)^2/2(\kappa^2+\kappa_g^2)\right\}} \\ &= (1-c_g)^{-1/2}\exp\left\{-\frac{1}{2}\frac{[c_g(x_g-\mu_g)+(1-c_g)\theta]^2}{c_g\kappa_g^2}+\frac{\theta^2}{2\kappa^2}\right\} \\ &\propto (1-c_g)^{-1/2}\exp\left\{-\frac{1}{2}T_g^2\right\}, \qquad (15)\end{aligned}$$

with the constant of proportionality being $\exp(\theta^2/2\kappa^2)$, and where

$$c_g = \frac{1}{\kappa_g^2}\left(\frac{1}{\kappa_g^2}+\frac{1}{\kappa^2}\right)^{-1} = \frac{1}{1+\kappa_g^2/\kappa^2}.$$

Similarly, for the other likelihood ratio we have

$$\frac{L_2(x_g)}{L_1(x_g)} = \exp\left\{-\frac{2(x_g-\mu_g)\theta}{\kappa^2+\kappa_g^2}\right\}. \qquad (16)$$

Suppose that $\theta > 0$, so that δ_{1g} is an indicator of inflated variance. Then, L_2/L_1 converges to zero as x_g increases to infinity and so, in the limit, δ_{1g} is solely a function of the ratio, L_0/L_1, and hence of the statistic, T_g. On the other hand, L_2/L_1 converges to infinity as x_g decreases to $-\infty$ so that δ_{1g} converges to zero. This makes sense since, in this case, it is highly unlikely that the gene is in the inflated variance non-null group. Parallel arguments can be made regarding δ_{2g}.

Note that $x_g-\mu_g$ is the observed difference between the log variances in the control and treatment groups for gene g (after adjusting for the covariate effects), and θ represents the expected difference if the gene has inflated variance under treatment (assuming $\theta > 0$). Thus, the numerator of the statistic, T_g, has the form of classical James-Stein shrinkage estimator of difference in the log variances, with the amount of shrinkage of the observed difference towards θ determined by the ratio of variances κ_g^2/κ^2.

5 Simulation Results

We compared the performance of the two estimation procedures in terms of power, accuracy, and false discovery rate with the 'one hypothesis at a time' approach, using the Brown and Forsythe (1974) median centered robust version of Levene's test

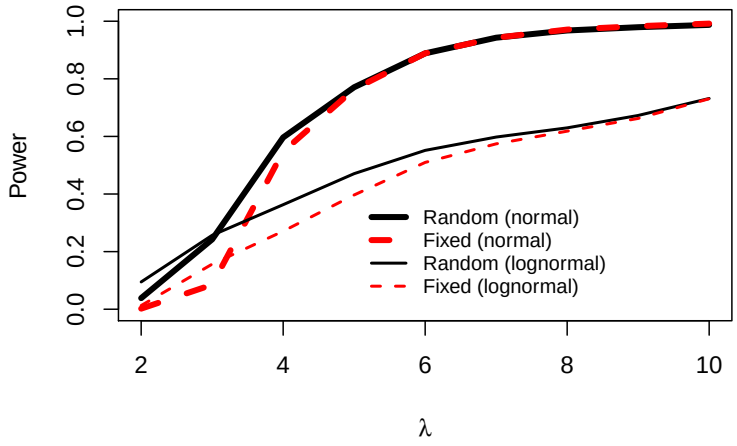

Figure 1: Power as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively.

(Levene, 1960). We chose the Levene test since previous studies have shown it to be relatively robust and powerful (Gastwirth et al., 2010). We also compared our approach with other well-known 'one at a time' methods, like Bartlett's test (Bartlett, 1937), although those comparisons are not reported here. The traditional methods that do not borrow strength across levels lack power, especially when the sample sizes are small. See Boos and Brownie (2004) and Boos and Brownie (1989) for comprehensive reviews of 'one at a time' methods and their power and robustness properties.

In our simulations we varied the sample sizes, ranging from $n_i = 2$ to $n_i = 30$, and allowed for the two groups to have different sample sizes. The inflation factor varied in the range $2 \leq \lambda \leq 10$, and we used a variety of underlying distributions, including normal, Cauchy, lognormal, and exponential, in order to assess robustness. Each simulation configuration was repeated 20 times, and the results are reported in terms of the average of the 20 experiments. The configurations reported in this paper involve sample sizes, $n_1 = 4$ and $n_2 = 7$, $G = 2000$ genes with a 10% inflated-variance subset ($p = 0.1$). The underlying distributions of the responses, y_{ijg}, are $N(0, 0.25)$, $LN(0, 0.25)$, and Cauchy distribution, with location and scale parameters equal to 0 and 0.1, respectively. The results reported here are representative of the wide range of simulation studies that we performed. The software used to perform this analysis is available from the authors. It will be released as an update to the `lemma` package (Bar and Schifano, 2010) shortly.

Figure 1 shows the power of the random factor (solid line) and the fixed factor (dashed line) approaches for two configurations; one, for normal data, $y_{null} \sim$

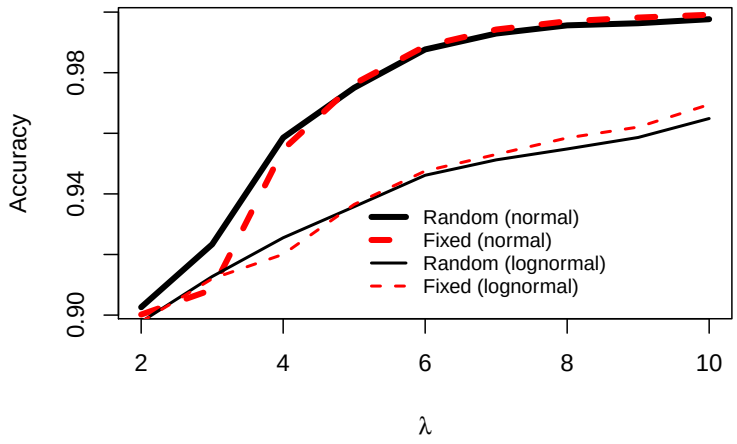

Figure 2: Accuracy as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively.

$N(0,0.25)$ (thick lines), and one for lognormal data, $y_{null} \sim LN(0,0.25)$ (thin lines). To generate these power plots we used the frequentist-type inference, and controlled for false discovery rate at the 5% level. The 'one hypothesis at a time' approach using the Levene test had zero discoveries for any λ (after applying the Benjamini-Hochberg adjustment). The 'random inflation factor' approach is more powerful than the 'fixed factor' procedure, in both configurations. As expected, as the true inflation factor increases both procedures become more powerful. Also, the power is higher when the underlying data are normally distributed.

Of course, in addition to power we would like the methods to have high level of accuracy (the total percentage of correct classifications, i.e., $100\times$(True Positive + True Negative)/G). Figure 2 shows (for normal and lognormal data) that both methods are quite accurate, and their accuracy increases as the inflation factor increases. In contrast, the conservative one-at-a-time approach, as well as the mean-based methods (not shown in the plot), yield approximately constant level of accuracy (in this case, 0.9, since by not rejecting **any** test, they correctly classify the null subset.)

ROC curves (of the average true positive rate versus the false positive rate) are given in Figure 3 for the normal, lognormal, and Cauchy data when the inflation factor is $\lambda = 4$. The three ROC plots are confined to a false positive rate of less than or equal to 0.2 since higher error rates than this would clearly be undesirable. In all cases the random factor model has the best performance. For example, when the data are normal, at a false positive level of 0.05 the average true positive rates are approximately 0.2, 0.5, and 0.65 for the (median-centered) Levene, fixed inflation

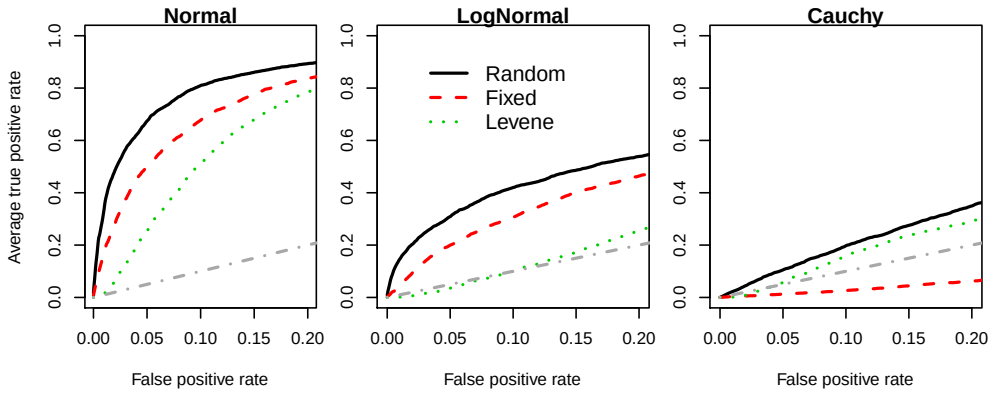

Figure 3: ROC curves, $\lambda = 4$. The solid and dashed lines correspond to the random and fixed factor models, respectively. The dotted lines correspond to the median centered robust version of Levene's test. 'Random classification' is represented by the dot-dashed line.

factor, and random factor methods, respectively.

When the data are normal or lognormal, both the random and fixed factor models are much better than the Brown and Forsythe (1974) median centered robust version of Levene's test. In particular, the middle plot shows that Levene's test is not at all robust to the normality assumption, as its ROC curve falls below the 'random classification' line (in grey). In contrast, under the lognormal data generation scheme, both the random and fixed factor models are quite robust. Furthermore, for all the simulated distributions the performance of our methods improves as the inflation factor increases, but the Levene method does not exhibit any improvement (not shown in the plot).

The fixed factor method performs very poorly with Cauchy data (right panel). In fact, it is even worse than the Levene method. An explanation is that the estimate of λ is not consistent because the mean of the Cauchy distribution does not exist, so the fixed factor model is clearly not appropriate in this extreme case. In contrast, the random factor model allows for variability in the distribution of the inflation factor, and is able to detect a reasonable number of the genes with differential variance while maintaining a low false positive rate.

The interpretation of the ROC plots requires care. It appears that for the normal data the 'one at a time' method has comparable performance to the two model-based approaches since, although the Levene-based ROC curve is below the other two model-based curves, it is well above the diagonal (in grey). However, it merely indicates that for a certain threshold of the p-values, the number of true positives exceed the number of false positives. In practice, the thresholds used to plot the ROC curve for the Levene test are much too high to be practical in real-life

applications, since, as we discussed above, the 'one at a time' method yields no discoveries at any reasonable FDR threshold when the number of tests is large.

6 Case Studies

We consider four different statistical applications to genetics and molecular biology to demonstrate the wide range of data sets to which our method can be applied. In all cases we find strong evidence that there is a subset of the data in which the variance in the treatment group is significantly higher or lower than in the control group but there is no significant difference between the means. The first case study involves a gene expression data set. The second deals with epigenetic data (methylation), while the third uses data from a brain imaging experiment (functional MRI data). The final example concerns metabolomics data.

The results in this section illustrate two things that are relevant to our previous derivations. First, we see that the observed distributions of the statistics r_g and x_g in the applications considered are very close to the ones in our model. In particular, the normal approximation of x_g appears to be very appropriate. Second, when the overall mean does not change due to the treatment, but the variance does, our method is able to detect it. In that sense, it complements the mean-based methods, which would (most likely) fail to detect the change in variance, unless it is coupled with a significant change in the mean response.

6.1 Microarray Data

Callow et al. (2000) used gene targeting in embryonic stem cells to produce mice lacking apolipoprotein A-1, a gene known to play a critical role in high density lipoprotein (HDL) cholesterol levels. In our analysis, we used the data and normalization method provided with the `limma` R package (Smyth, 2005), which consists of 5,548 ESTs, from eight control (wild type "black six") mice and eight "knockout" (lacking ApoA1) mice. Common reference RNA was obtained by pooling RNA from the control mice, and was used to perform expression profiling for all 16 mice. Using the `lemma` package (Bar and Schifano, 2010), which is designed to detect genes that are differentially expressed, 9 genes are detected (with a 0.2 posterior probability threshold) including the ApoA1 gene and others closely related to it. The same set of the top eight genes were also identified as non-null (among others) when using other (mean-based) packages like `limma` and `locfdr` (Efron et al., 2008). These genes were confirmed to be differentially expressed in the knockout versus the control line by an independent assay.

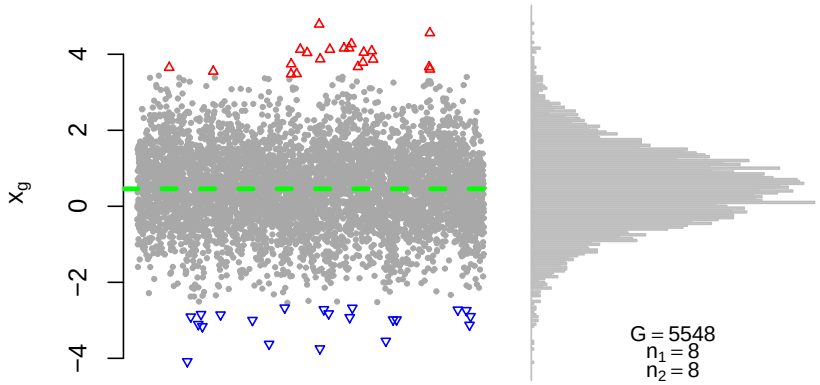

Figure 4: The distribution of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$, case study 1: the Apo-A1 data set. Number of genes $G = 5548$, sample size $n_1 = n_2 = 8$, FDR threshold=0.05.

Applying the method in this paper while controlling the false discovery rate at 5% we find 21 genes in which the variance in the treatment group was significantly higher than in the control, and 21 genes in which the variance was significantly smaller in the treatment group. Most of these genes had very small mean-response difference (defined as $d_g = \bar{y}_{2\cdot g} - \bar{y}_{1\cdot g}$) and were not detected by any mean-based method, or by 'one at a time' test for unequal variance.

Figure 4 shows the distribution of the statistics $\{x_g\}$. The scatter plot on the left shows the genes with significantly higher and lower variance, marked by upper red or lower blue triangles, respectively. The scatter plot and the histogram (right) show that the normal approximation fits the distribution of x_g very well. The green dashed line represents the overall mean of x_g (which, in our previous notation, we referred to as $\log(\tau)$).

We investigated the functional status of the genes that had deflated and inflated variances using the National Institute of Health Gene tool and Genomenet (`http://www.ncbi.nlm.nih.gov/gene` and `http://www.genome.jp/`, respectively). It turns out that the inflated variance genes mostly have to do with cell signaling, while the deflated variance genes seem more related to tighter regulation of a lipid metabolism gene network. Given that the gene of primary focus in the study, ApoA1, encodes apolipoprotein A-I, which is the major protein component of high density lipoprotein (HDL) in plasma these results are biologically plausible.

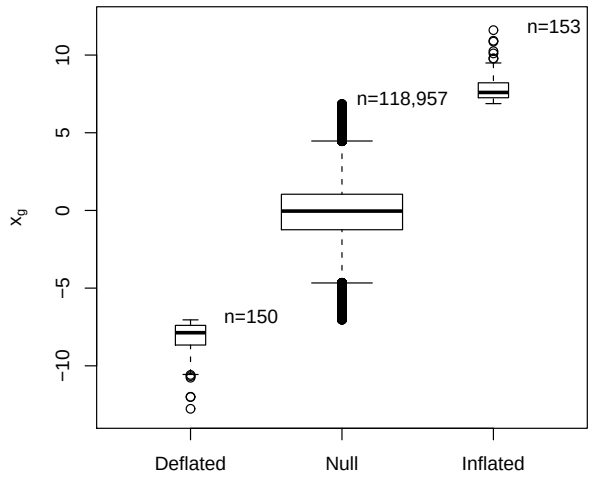

Figure 5: Case study 2: methylation data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of genes $G = 119,260$, sample size $n_1 = n_2 = 3$, FDR threshold=0.05.

6.2 Methylation Data

DNA methylation plays an important role in regulation of gene expression. Recent studies have shown that hyper- or hypomethylation are associated with cancer (either as a causal effect or as an early indicator of the disease). In the following analysis, we used an unpublished data set with 119,260 genes, and three subjects in each group. Using the mean-based approach (Bar et al., 2010) we did not find any significantly hyper or hypomethylated genes. However, applying the methods developed in this paper we found a total of 153 genes with inflated methylation, and 150 with deflated methylation (at the FDR level of 5%). In contrast, traditional 'one at at time' methods yield no discoveries, after accounting for multiple testing.

The observed mean differences $\{d_g\}$ are rather small, but the observed log-ratio between the mean squared errors are very large (in absolute value) for some genes. Figure 5 shows the boxplots of the three mixture components. The distribution of x_g in the null component is approximately normal with mean 0, and the significant genes have $|x_g| > 7$. Recall that x_g is on the logarithmic scale, so for the significant genes this corresponds to at least four orders of magnitudes in the ratio between the mean squared errors between the two groups.

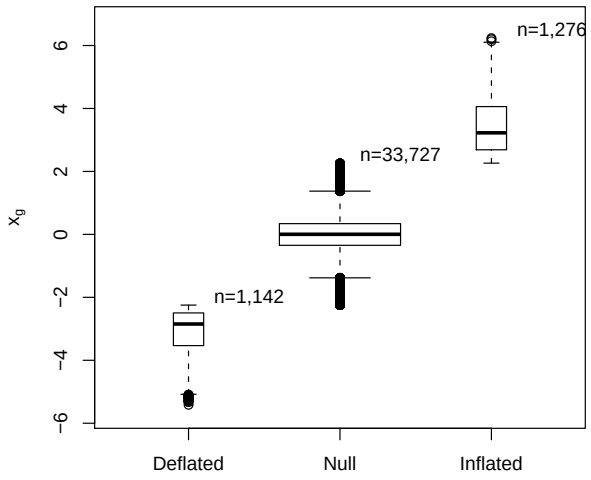

Figure 6: Case study 3: fMRI data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of voxels $G = 36,145$, sample size $n_1 = 29, n_2 = 22$, FDR threshold=0.05.

6.3 fMRI Data

Functional magnetic resonance imaging (fMRI) is used to measure the change in blood flow in the brain during certain neural or cognitive activity. In this example we use data from a Pavlovian-type experiment, in which both groups were shown a visual cue, but for the treatment group it was immediately followed by an auditory signal (Soliman et al., 2010). One of the goals of the experiment was to test whether after several training cycles there is a difference in the response to the visual cue between the two groups, and if so, in which region of the brain. According to the Pavlovian paradigm, it is expected that once trained, the treated subjects will respond to the visual cue as if they receive the auditory cue. For more details about the experiment, see the 'Supporting Online Material' document in Soliman et al. (2010).

Again, no voxels were found to have significantly different mean levels of response when using mean-based methods. However, we do find many voxels which exhibit significantly different levels of variability. Figure 6 shows the boxplots of the three mixture components that our method identified from a total of 36,145 voxels. A total of 1,276 voxels had a significantly increased variance in the treatment group, and 1,142 voxels had a significantly decreased variance in the treatment group.

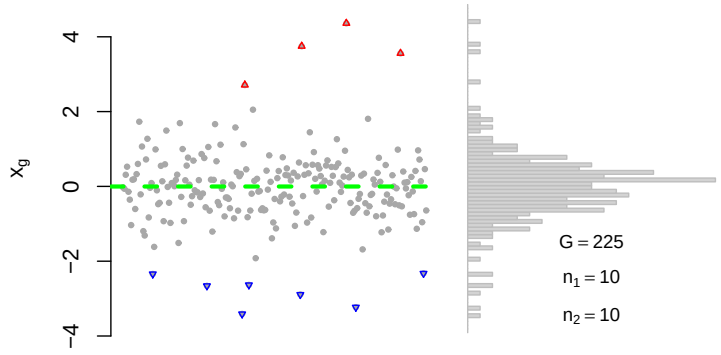

Figure 7: The distribution of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$, case study 4: metabolomics data set. Number of metabolites $G = 225$, sample size $n_1 = n_2 = 10$, FDR threshold=0.05.

6.4 Metabolomics Data

Our final example uses data from the area of metabolomics. In recent years the study of chemical processes involving metabolites has become a popular complement to gene-expression analysis. Metabolites are the product of cellular processes, and they include several pathways, like amino acid (e.g., creatine, glutamine), lipid (e.g., choline, 2-hydroxyglutarate), and energy (citrate, pyrophosphate), just to name a few. Cancer researchers noted that there are differences in cellular metabolism between normal and cancer cells (DeBerardinis et al., 2007).

In this (unpublished) experiment, two groups of pregnant women were treated with two different levels of choline. The levels of nearly 250 metabolites were measured during the first and the twelfth weeks of the pregnancy. Here, we analyze the effect of the treatment on metabolite levels after 12 weeks (taking week 0 as the baseline for each woman). Once again, testing for differences in mean response levels between the groups yields no discoveries. However, with our method we find four metabolites whose variance increased significantly due to the treatment, and seven whose variance decreased (see Figure 7).

7 Conclusions

We have developed a new model and an estimation procedure (based on the EM algorithm) for parallel testing for inequality of variances. The model borrows strength across the entire data, resulting in increased power and accuracy, while maintaining

a low false discovery rate. Simulations show that the method performs well even when the number of tests is very large and the sample sizes are small, and that it is quite robust to deviations from normality. Our analysis of four different data sets shows that the model assumptions are realistic, that the method is broadly applicable, and that it complements methods that test for differences in means. In future work we hope to develop a bivariate procedure for simultaneous testing of means or variances.

References

Bar, H., J. Booth, E. D. Schifano, and M. T. Wells (2010): "Laplace approximated em microarray analysis: An empirical bayes approach for comparative microarray experiments," *Statistical Science*, 25, 388–407.

Bar, H. and E. Schifano (2010): *lemma: Laplace approximated EM Microarray Analysis*, URL `http://CRAN.R-project.org/package=lemma`, r package version 1.3-1.

Bartlett, M. S. (1937): "Properties of sufficiency and statistical tests," *Proceedings of The Royal Society of London. Series A, Mathematical and Physical Sciences (1934-1990)*, 160, 268–282.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate-a practical and powerful approach to multiple testing," *Journal of The Royal Statistical Society Series B*, 57, 499–517.

Boos, D. D. and C. Brownie (1989): "Bootstrap Methods for Testing Homogeneity of Variances," *Technometrics*, 31, 69–82, URL `http://dx.doi.org/10.2307/1270366`.

Boos, D. D. and C. Brownie (2004): "Comparing variances and other measures of dispersion," *Statistical Science*, 19, 571–578.

Box, G. E. P. (1953): "Non-normality and tests on variances," *Biometrika*, 40, 318–335.

Brown, M. and A. Forsythe (1974): "Robust tests for equality of variances," *Journal of the American Statistical Association*, 69, 364–367.

Cai, L., N. Friedman, and X. S. Xie (2006): "Stochastic protein expression in individual cells at the single molecule level," *Nature*, 440, 358–362, URL `http://dx.doi.org/10.1038/nature04599`.

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

Colman-Lerner, A., A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C. Gustavo Pesce, and R. Brent (2005): "Regulated cell-to-cell variation in a cell-fate

decision system," *Nature*, 437, 699–706, URL http://dx.doi.org/10.1038/nature03998.

DeBerardinis, R. J., A. Mancuso, E. Daikhin, I. Nissim, M. Yudkoff, S. Wehrli, and C. B. Thompson (2007): "Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis." *Proceedings of the National Academy of Sciences of the United States of America*, 104, 19345–19350, URL http://dx.doi.org/10.1073/pnas.0709747104.

Efron, B. (2008): "Microarrays, empirical bayes and the two groups model," *Statistical Science*, 23, 1–22.

Efron, B. and C. Morris (1975): "Data Analysis Using Stein's Estimator and its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Efron, B., B. B. Turnbull, and B. Narasimhan (2008): *locfdr: Computes local false discovery rates*, R package version 1.1-6.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau (2010): "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, 11, 446–450, URL http://dx.doi.org/10.1038/nrg2809.

Feinberg, A. P. and R. A. Irizarry (2010): "Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease," *Proceedings of the National Academy of Sciences*, 107, 1757–1764, URL http://dx.doi.org/10.1073/pnas.0906183107.

Gastwirth, J. L., Y. R. Gel, and W. Miao (2010): "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice," *Statistical Science*, 24, 343–360.

Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg (2011): "Increased methylation variation in epigenetic domains across cancer types," *Nature Genetics*, 43, 768–775, URL http://dx.doi.org/10.1038/ng.865.

Ho, J. W., M. Stefani, C. G. dos Remedios, and M. A. Charleston (2008): "Differential variability analysis of gene expression and its application to human diseases," *Bioinformatics*, 24, i390–i398, URL http://dl.acm.org/citation.cfm?id=1388083.1388131.

Hwang, J. T. G. and P. Liu (2010): "Optimal tests shrinking both means and variances applicable to microarray data analysis." *Statistical Applications in Genetics and Molecular Biology*, 9, Article36.

James, W. and C. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 361–379.

Levene, H. (1960): "Robust tests for equality of variances," *In Contributions to Probability and Statistics*, 278–292.

Levsky, J. M., S. M. Shenoy, R. C. Pezo, and R. H. Singer (2002): "Single-cell gene expression profiling." *Science (New York, N.Y.)*, 297, 836–840, URL `http://dx.doi.org/10.1126/science.1072241`.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher (2009): "Finding the missing heritability of complex diseases," *Nature*, 461, 747–753, URL `http://dx.doi.org/10.1038/nature08494`.

Mar, J. C., N. A. Matigian, A. Mackay-Sim, G. D. Mellick, C. M. Sue, P. A. Silburn, J. J. McGrath, J. Quackenbush, and C. A. Wells (2011): "Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease," *PLoS Genetics*, 7.

Mar, J. C., R. Rubio, and J. Quackenbush (2006): "Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples," *Genome Biology*, 7, R119+, URL `http://dx.doi.org/10.1186/gb-2006-7-12-r119`.

Marko, N. F., J. Quackenbush, and R. J. Weil (2011): "Why is there a lack of consensus on molecular subgroups of glioblastoma? understanding the nature of biological and statistical variability in glioblastoma expression data," *PLoS ONE*, 6.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden (2002): "Regulation of noise in the expression of a single gene." *Nature Genetics*, 31, 69–73, URL `http://dx.doi.org/10.1038/ng869`.

Ravasi, T., C. Wells, A. Forest, D. Underhill, B. Wainwright, A. Aderem, S. Grimmond, and D. Hume (2002): "Generation of diversity in the innate immune system: macrophage heterogeneity arises from gene-autonomous transcriptional probability of individual inducible genes." *Journal of Immunology*, 168, 44–50.

Smyth, G. K. (2004): "Linear Models for Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*, 3, article 2.

Smyth, G. K. (2005): *Limma: Linear Models for Microarray Data*, New York: Springer, 397–420.

Soliman, F., C. E. Glatt, K. G. Bath, L. Levita, R. M. Jones, S. S. Pattwell, D. Jing, N. Tottenham, D. Amso, L. H. Somerville, H. U. Voss, G. Glover, D. J. Ballon, C. Liston, T. Teslovich, T. V. Kempen, F. S. Lee, and B. J. Casey (2010): "A genetic variant bdnf polymorphism alters extinction learning in both mouse and human," *Science*, 327, 863–866.