Unequal group variances in microarray data analyses

Meaza Demissie¹, Barbara Mascialino², Stefano Calza³ and Yudi Pawitan²,*
¹Department of Statistics, University of Örebro, Sweden, ²Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and ³Department of Biomedical Sciences and Biotechnology, University of Brescia, Italy

Received on January 21, 2008; revised on February 23, 2008; accepted on March 12, 2008
Advance Access publication March 14, 2008
Associate Editor: Martin Bishop

ABSTRACT

Motivation: In searching for differentially expressed (DE) genes in microarray data, we often observe a fraction of the genes to have unequal variability between groups. This is not an issue in large samples, where a valid test exists that uses individual variances separately. The problem arises in the small-sample setting, where the approximately valid Welch test lacks sensitivity, while the more sensitive moderated t-test assumes equal variance.

Methods: We introduce a moderated Welch test (MWT) that allows unequal variance between groups. It is based on (i) weighting of pooled and unpooled standard errors and (ii) improved estimation of the gene-level variance that exploits the information from across the genes.

Results: When a non-trivial proportion of genes has unequal variability, false discovery rate (FDR) estimates based on the standard t and moderated t-tests are often too optimistic, while the standard Welch test has low sensitivity. The MWT is shown to (i) perform better than the standard t, the standard Welch and the moderated t-tests when the variances are unequal between groups and (ii) perform similarly to the moderated t, and better than the standard t and Welch tests when the group variances are equal. These results mean that MWT is more reliable than other existing tests over wider range of data conditions.

Availability: R package to perform MWT is available at http://www.meb.ki.se/~yudpaw
Contact: yudi.pawitan@ki.se
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Statistical analyses of microarray data have reached a certain consensus, for example (i) assessment based on false discovery rates (FDR) is more relevant than one based on p-values (e.g. Storey and Tibshirani, 2003) and (ii) in small samples the standard t-test can perform poorly, so it should be replaced by some modified versions (e.g. Baldi and Long, 2001; Efron et al., 2001; Lönnstedt and Speed, 2002; Ploner et al., 2006; Smyth, 2004). This article will focus on one issue for which there is yet no consensus: how to deal with unequal variability between groups. By ‘unequal variance problem’, we mean that a non-trivial proportion of the genes have unequal group variances. For clarity and simplicity, we will limit our discussion to the most common problem of finding differentially expressed (DE) genes between two biological conditions. We show that unequal group variance is a common problem in practice, and, in this setting, standard FDR estimates based on tests that assume equal variance can give misleading optimistic bias. The standard remedy using the Welch test is valid in large samples, but lacking in sensitivity in small samples. We propose a moderated version of the Welch test, and show that it works well in small samples.

1.1 Notation and standard tests

We start with the standard tests as a way of introducing the notation. Let n₁ and n₂ be the number of arrays for the two groups, with each array containing probes for m genes. For gene g, we observe a vector of expression values yᵢᵍ of length n₁ + n₂, which consists of the observations yᵢ in the first group, and yᵢ₂ in the second group. We define the group means and SDs as usual, but for convenience we drop the subscript g. Denote the standard t-statistic by

\[ t = \frac{\bar{y}_1 - \bar{y}_2}{s_e}, \]

using pooled (squared) standard error

\[ s_e^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right). \]

To allow for unequal variances, one might consider the so-called Welch’s t-statistic

\[ W = \frac{\bar{y}_1 - \bar{y}_2}{s_w}, \]

with the unpooled standard error

\[ s_w^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}. \]

In large samples, W has an approximate standard normal distribution regardless of the variances, so no problem arises. However, in small samples the normal approximation does...
not apply. Instead, under the null, $W$ has an approximate $t$ distribution with estimated degrees of freedom given by

$$d' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

In microarray data analysis, the standard $t$-test is known to have poor sensitivity (see e.g. Ploner et al., 2006). This property can be expected to occur also with the Welch version. The moderated $t$-statistic (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004), introduced to improve on the performance of the standard $t$-test, has the form

$$t_m = \frac{\bar{y}_1 - \bar{y}_2}{\hat{s}_p},$$

where

$$\hat{s}_p = \frac{s^2(1/n_1 + 1/n_2)}{d_0},$$

$$s^2 = \frac{(d_0s_1^2 + d_2s_2^2)}{(d_0 + d)},$$

$$d = n_1 + n_2 - 2,$$

and $d_0$ and $s_0^2$ are hyper-parameters to be estimated from the data. There are other improvements of the $t$-statistic, e.g. Efron’s (Efron et al., 2001) or SAM (Tusher et al., 2001), or the multidimensional FDR (Ploner et al., 2006). However, these forms have unknown null distribution, so they require large-enough samples to allow a permutation test, while in this article our objective is to have a procedure that can work in small-sample problems where the permutation-based null distribution is not available. One key advantage of $t_m$ is that under the null it has the standard $t$ distribution with $d_0 + d$ degrees of freedom, so no permutation computation is needed for inference and the procedure can be applied to small-sample comparisons.

In principle, Ploner et al.’s multidimensional FDR idea (Ploner et al., 2006) also works for unequal variance case, but, again, the sample size must be large enough to generate permutation-based null distribution.

### 1.2 Unequal-variance problem

Unequal group variances occur frequently in practice. Figure 1 shows four different examples of the histogram of $p$-values from $F$-tests for equal variance; see Section 2.1 for the data description. The $F$-test is the most commonly used test of equal variance, but in principle any test of equal variance is applicable; for example, one might consider a test that is less sensitive to normal assumption. If the group variances are equal, the distributions should be uniform. In fact each histogram shows an excess of small $p$-values, indicating a sizable proportion of genes with unequal group variances. Using formula (6), we estimate the proportion of genes with unequal group variances as 0.10, 0.21, 0.38 and 0.20 for panels (a) to (d) respectively.

What is the effect of unequal variance on the search for DE genes? Figure 2a shows a simulation result under the null setup of no DE genes, comparing $n_1 = 3$ versus $n_2 = 3$ arrays of $m = 10000$ genes. The correct $F$-test is equal to one, but the estimated $FDR$s based on the standard $t$ and $t_m$ are clearly biased downwards, leading to declaration of DE genes with misleadingly small $FDR$ estimate. What is surprising is that the problem is already visible for a balanced sample, where these tests are known to be robust against unequal variance. As we show later, the bias is worse for unbalanced samples. The Welch test $W$ performs as expected with little bias.

Does this mean that one should always use $W$? In large samples, $W$ converges to the standard normal regardless of equality of the variances, so it is more robust than the standard $t$. However, in small samples, $W$ suffers the same lack of sensitivity as the standard $t$. Figure 2b shows a simulation result comparing $n_1 = 3$ versus $n_2 = 3$ arrays, where there are truly DE genes. The true $FDR$s based on $t$ (solid) and $W$ (dotted) coincide, and these are substantially larger than the $FDR$ achieved by the moderated statistic $t_m$.

### 1.3 Moderated Welch test

Suppose, as we expect, there is a subset of genes where the group variances are equal and others where the variances are not equal. Had we known which is the case for each gene,
we would use the pooled formula when the group variances are equal, and unpooled otherwise, i.e.

\[ se^2 = I(\sigma_1^2 = \sigma_2^2) \hat{se}_1^2 + I(\sigma_1^2 \neq \sigma_2^2) \hat{se}_u^2, \]

where the indicator \( I(\cdot) = 1 \) if the condition is true, and zero otherwise. In practice we do not know whether \( \sigma_1^2 = \sigma_2^2 \), but we can investigate it empirically. This suggests a probability weighting:

\[ \hat{se}_w = w \hat{se}_p^2 + (1-w)\hat{se}_u^2, \]

where \( w \) is the (posterior) probability that the variances are equal given the data. If we use the F-test of equal variance, we can condition on the extremeness of the observed statistic: \( F > F_{\text{obs}} \). The probability \( P(\sigma_1^2 = \sigma_2^2 | F > F_{\text{obs}}) \) is in fact given by the FDR for testing the equality of variance, so it can be estimated from the data using standard FDR computations.

\[ \Gamma(x) = \frac{1}{2} \left( \frac{\pi}{x} \right)^{1/2} \Gamma \left( \frac{1}{2} \right) \left( x^2 - \frac{1}{2} \right)^{-1/2} \]

where \( \Gamma \) is the gamma function.

Intuitively, if two observed variances are very different, it will be easier to get a much higher weight. We then approximate the distribution of \( \hat{se}_w^2 \) as a scaled \( \chi^2 \) with degrees of freedom given by

\[ d_w = wd + (1-w)df, \]

where \( d \) and \( df \) are the degrees of freedom of the pooled and unpooled standard errors, respectively.

It is clear that probability weighting alone is not sufficient to get a sensitive procedure. For example, if all the group variances are equal, then the weighted formula would yield the pooled standard errors, and we would end up with the standard t-test. We now propose a moderated Welch test (MWT), whose form is motivated by the moderated \( t_m \) statistic:

\[ W_m = \frac{\bar{y}_1 - \bar{y}_2}{\hat{se}_m}, \]

based on the moderated standard error

\[ \hat{se}_m^2 = d_0 \hat{se}_0^2 + d_u \hat{se}_u^2, \]

where \( d_0 \) and \( \hat{se}_0^2 \) are again hyper-parameters to be estimated from the data. When the variances are equal, \( W_m \) reduces to the moderated \( t_m \) statistic, but we expect to reduce bias if the variances are unequal.

1.4 Summary

To summarize our findings, when measurement variability is unequal between groups: (i) the use of \( t \) or \( t_m \) leads to more false positives than declared by the standard FDR estimates, (ii) the standard Welch test is unbiased, but lacks sensitivity, (iii) MWT deals with the unequal-variance problem without the loss of power suffered by the standard Welch test, and finally, (iv) when group variances are in fact equal, MWT and \( t_m \) perform similarly. These results mean that MWT is a more robust test to use than the standard \( t \), \( t_m \) or \( W \) tests over a wide range of data conditions.

2 METHODOLOGY

2.1 Datasets

To exhibit the common occurrence of unequal group variances (as shown in Figure 1) we used these datasets:

- The BRCA dataset (Hedenfalk et al., 2001), collected from patients with hereditary breast cancer, who had mutations either of the BRCA1 (\( n_1 = 7 \)) or the BRCA2 gene (\( n_2 = 8 \)). After quality control, we are left with 3170 genes.
- A dataset of 240 cases of diffuse large B-cell lymphoma (Rosenwald et al., 2002) with 7399 genes, where we compare \( n_1 = 102 \) survivors and \( n_2 = 138 \) non-survivors.
- A dataset of 159 population-based breast cancer cases (Pawitan et al., 2005a,b); we used the Affymetrix U133A chips and filtered out probes with <50% presence across the samples, leaving us with 11295 genes for analysis. We compared the group variances of two key variables: (a) estrogen receptor (ER) status (\( n_1 = 130 \) positive and \( n_2 = 29 \) negative) and (b) survival status at 5 years (\( n_1 = 121 \) alive and \( n_2 = 38 \) dead). Note the unbalanced sample sizes in these examples.

For procedure comparisons, we also used the so-called ‘Golden-Spike’ data (Choe et al., 2005). It is a dataset of 3860 RNA species, where 100–200 RNAs were spiked in at fold-change (FC) level ranging from 1.2- to 4-fold, while a set of 2551 RNA species was spiked-in at a constant (FC = 1) level. Data were designed as a two-group comparison, spike-in (S) versus control (C) (\( n = 3 \) each). There is a total of 14010 probesets, and in our comparisons we define as true DE those probes that have fold-change at least 3, yielding 454 probes.

2.2 Mean and variance models

To allow small-sample inferences, we will assume a parametric model as follows. Given a gene-wise mean \( \mu_{gi} \) and variance \( \sigma_{gi}^2 \) from group \( i \), the log-expression value of the \( j \)th sample is

\[ y_{gji} \sim N(\mu_{gi}, \sigma_{gi}^2), \quad j = 1, \ldots, n_i. \]

To understand the variance pattern, we first analyze the observed gene-wise variances from the BRCA data and show the results in Figure 3. Some correlation (Pearson’s \( r = 0.4 \)) does exist, indicating a common element between the two groups, but the marginal distributions are similar. Similar patterns are observed in other datasets, with the correlation increasing with sample size, as expected since there is less sampling noise.

The group variances \( \sigma_{g1}^2 \) and \( \sigma_{g2}^2 \) vary across genes, so it is convenient to model them as random. In view of the pattern in Figure 3, we consider the model

\[ \log \sigma_{gj}^2 = \log \sigma_{gi}^2 + \epsilon_{gj}, \]

where \( \sigma_{gi}^2 \) is the common variance between the two groups, and \( \epsilon_{gj} \) is an independent perturbation term. The term \( \sigma_{gi}^2 \) will

![Fig. 3. Analysis of gene variances of the BRCA data: (a) Scatter plot of the variances from BRCA1 versus BRCA2; (b) The marginal density of the variances from BRCA1 (solid) and BRCA2 (dotted).](http://bioinformatics.oxfordjournals.org/)
account for the correlation in Figure 3a, and \( \epsilon_{gi} \) for the unequal variances in Figure 1. Note that \( \epsilon_{gi} = 0 \) for genes with equal group variance.

For explicit distribution theory, one option is to allow separate models for the two variances, but we have found that this approach did not work well. Instead we consider the following: first approximate the conditional distribution of the weighted standard error \( \text{se}_{wg}^2 \) as (a scaled) \( \chi^2 \) with parameter \( \nu_k^g \) and \( d_k^g \) degrees of freedom. Then use the inverse \( \chi^2 \) distribution to model \( \nu_k^g \) as

\[
\frac{1}{\nu_k^g} \sim \frac{1}{d_k^g} \chi^2_{d_k^g},
\]

where \( d_k^g \) is the degrees of freedom and \( d_k^g \nu_k^g \) the scale parameter. Then, the posterior distribution of \( \nu_k^g \) given \( \text{se}_{wg}^2 \) is inverse \( \chi^2 \) with \( (d_k^g + d_k^g + 2) \) and scale \( d_k^g \nu_k^g + d_k^g \nu_k^g \). So the posterior mean is

\[
\text{se}_{wg}^2 = \frac{\nu_k^g \text{se}_{wg}^2}{d_k^g + d_k^g}.
\]

which is the standard error needed for the mean difference, leading to the moderated test \( W_m \) in (2). If the group variances are equal, this approach reduces to the standard \( t_m \) statistic.

### 2.3 Estimation of \( d_0 \) and \( s_0^2 \)

The hyper-parameters \( d_0 \) and \( s_0^2 \) reflect the variation of \( \text{se}_{wg}^2 \) across genes; because there is typically a large number of genes, the method-of-moments estimates for these parameters are adequate. Adapting Smyth (2004), we base our estimates on

\[
z_g = \log \text{se}_{wg}^2,
\]

which is roughly normal with mean and variance

\[
E(z_g) = \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d)
\]

\[
\text{var}(z_g) = \psi(d/2) + \psi(d_0/2),
\]

where \( \psi \) and \( \psi' \) are the digamma and trigamma functions, and \( d \) is the average of \( d_0 \). Using the sample versions of these moments, from the second equation we get an estimate of \( d_0 \); then we can compute \( s_0^2 \) from the first equation.

### 2.4 Simulation

We perform simulations with \( m = 10000 \) genes per array with two independent groups according to models (3) and (4). The common variance \( \sigma_g^2 \) is generated from an inverse \( \chi^2 \) distribution with unit scale and 10 degrees of freedom. We select a fraction \( \pi_{gi} \) of the genes to have common group variance, and the rest to have unequal variance. Except when stated otherwise, we use \( \pi_{gi} = 0.7 \) throughout. For genes with unequal variances, \( \epsilon_{gi} \) in (4) is an independent sample from \( N(0, 1) \).

We simulate null and non-null cases under various scenarios as described in Sections 4.2 and 4.3. In the non-null case, we use a proportion of truly non-DE genes \( \pi_{gi} = 0.9 \); for each DE gene, we set \( \mu_{gi} = 0 \) and \( \mu_{gi} \sim N(0, \sigma_g^2) \).

For each scenario we use 100 replications (for a total of \( 10^6 \) genes). In order to evaluate the performance of the different procedures we compare the global FDR associated with the \( k \) top ranking genes and draw the resulting true FDR curve as a function of the proportion of genes declared DE. For each procedure, the genes are first ranked according to the test statistics, then the FDR is computed from the known DE status as a proportion of false positives. That is,

\[
\text{FDR}(k) = \text{Average} \left\{ \frac{\text{Number of non-DEs among top } k \text{ genes}}{k} \right\}
\]

where the average is taken over 100 replications.

### 2.5 FDR estimation

The FDR in the previous subsection is the true FDR achieved by a test procedure, which is an unknown parameter. In reality, this FDR needs to be estimated. We will use the usual estimator, for the \( k \) top genes,

\[
\hat{\text{FDR}}(k) = m \pi_0 p_{\hat{k}/k},
\]

where \( p_{\hat{k}/k} \) is the \( k \)th highest \( p \)-value; monotonicity is then imposed by applying a cumulative minimum over \( k, \ldots, m \). The proportion \( \pi_0 \) of non-DE genes is estimated (e.g. Storey and Tibshirani, 2003) by

\[
\hat{\pi}_0 = \frac{\text{Number of } p \text{-values } > \lambda}{m(1 - \lambda)}
\]

In our computations we use a cutoff value \( \lambda = 0.7 \), but the results are similar for values between 0.5 and 0.9.

The computation of \( p \)-values is straightforward, since all null distributions are known. If the sample sizes are large enough, one might consider the permutation method to compute the \( p \)-values. In our computations we assume parametric models, since we wish to use the procedures in small samples.

### 2.6 FDR weights

The FDR weight \( w \) in (1) is computed as the estimated FDR (5) for group-variance comparisons. Under model (3) and under the null hypothesis of equal group variance, we have

\[
\frac{s_1^2}{s_2^2} \sim F_{(1, \nu_1 - 1), (\nu_2 - 1)},
\]

so the \( p \)-value can be computed easily. The parameter \( \pi_{i0} \) is similarly estimated using (6). In summary, for a gene \( g \) with sample variances \( s_1^2 \) and \( s_2^2 \), we obtain an \( F \)-statistic, \( p \)-value \( p_g \) and weight \( w_g = \text{FDR}(p_g) \).

### 3 OTHER METHODS FOR COMPARISONS

Three additional methods for dealing with small samples and unequal variances were compared to MWT, namely local-pooled-error (LPE), weighted analysis of microarray experiments (WAME) and BGmix:

- The LPE test for DE genes (Jain et al., 2003) is based on a shrinken estimate of the within-group error variance.
- The WAME (Sjögren et al., 2007) assumes the arrays are dependent, with a certain variance–covariance matrix that accounts for heteroscedasticity and correlation.
- BGmix (Lewin et al., 2007) is a method for identifying DE genes based on a fully Bayesian mixture model. The model is formulated as a three-component mixture, describing genes that are under-, over- and non-DE.

More detailed descriptions are given in the Supplementary Report II. All these methods are available as R packages, and in all of our analyses we use their default settings. We also considered the method by Hu and Wright (2007) as implemented in the code available at the authors’
4 RESULTS

4.1 Completing Figure 2
Returning first to Figure 2, we now show the performance of MWT compared to the other tests. In this simulation we use $\pi_{0} = 0$, i.e. all genes have unequal group variances, but because of the small sample size ($n_1 = n_2 = 3$), most of the variance inequalities are too small to detect; in fact the estimated $\pi_{0}$ using formula (6) is around 0.7. In the null case—Figure 4a—the MWT is much less biased than the t or $t_m$ tests, while in the non-null case (b) MWT achieves the same performance as the moderated $t_m$.

4.2 Null case
It is known that the standard t-test is robust against unequal group variance, especially when the sample sizes are equal. Therefore, to investigate the effect of unequal variance on FDR estimates, we will consider balanced and unbalanced situations. Furthermore, the distribution of the Welch statistic is only approximate, so we expect a different performance in small and large samples. We set the proportion of genes with equal group variance to $\pi_{0} = 0.7$ throughout. Hence we consider the following four scenarios:

(a) Balanced small samples: $n_1 = 3$ and $n_2 = 3$.
(b) Unbalanced small samples: $n_1 = 3$ and $n_2 = 9$.
(c) Balanced larger samples: $n_1 = 10$ and $n_2 = 10$.
(d) Unbalanced larger samples: $n_1 = 10$ and $n_2 = 30$.

We first consider the null situation where there are no DE genes, so the true/correct FDR is equal to one for all procedures. Figure 5a shows that in a balanced small sample case, there is only a small bias in the FDR estimates. The bias becomes significantly worse when the data are unbalanced; see Figure 5b. Here all procedures show some bias, with MWT giving the smallest.

For balanced larger samples in panel (c), all procedures have very little bias. However, when the data are unbalanced (Figure 5d), only the Welch test and MWT are reliable. The standard t and moderated $t_m$ have a bias that persists in larger samples.

The figure for the comparison with LPE, WAME and BGmix is given the Supplementary Report II. In small samples, both balanced and unbalanced, LPE and WAME have little bias, with LPE having less bias. With bigger samples all methods have similar bias. In all four scenarios, BGmix suffers consistently worse bias than the others.

4.3 Non-null case
We now consider the case when there are truly DE genes. To reduce the number of plots, only two scenarios will be shown here (the other two are given in the Supplementary Report I):

- Balanced small samples: $n_1 = 3$ and $n_2 = 3$.
- Unbalanced larger samples: $n_1 = 10$ and $n_2 = 30$.

We will highlight two key issues: (i) different procedures have different sensitivity, which can be compared using the corresponding true FDRs (in contrast to the null case, where all procedures have the same true FDR, which is equal to one) and (ii) different procedures have different biases. Figure 6a shows that the FDRs of the MWT and moderated $t_m$ tests are lower than the FDRs of the standard t and Welch tests. This means that MWT mirrors the known improvement in terms of sensitivity of moderated $t_m$ over the standard t-tests. However, Figure 6b shows that the FDR estimates based on the moderated $t_m$ and MWT tests tend to be slightly biased. The standard t-test produces the worst negative bias, while the Welch test will give highly conservative estimates.

In larger unbalanced situation—Figure 6c—the true FDRs of the Welch and MWT are slightly better than the FDRs of the standard t and Welch tests. This means that MWT mirrors the known improvement in terms of sensitivity of moderated $t_m$ over the standard t-tests. However, Figure 6b shows that the FDR estimates based on the moderated $t_m$ and MWT tests tend to be slightly biased. The standard t-test produces the worst negative bias, while the Welch test will give highly conservative estimates.

Figures 7a and c show that WAME and MWT have similar true FDR both in small and larger samples, while LPE clearly...
produces higher FDR. However, both WAME and LPE have quite a large bias in the FDR estimation, both in small and big samples (Figs 7b and d). BGmix has the worst performance in both scenarios with a higher true FDR, and a consistently large negative bias.

Finally, in the Supplementary Report I we show that MWT performs similarly to the moderated $t_m$ when the group variances are in fact the same.

4.4 Golden-Spike data

Figure 8a shows the histogram of $p$-values for the $F$-test of equal variance of the spiked ($n = 3$) versus control ($n = 3$) groups of the Golden-Spike data. There is some indication of unequal variance, but it is not substantial. Furthermore, the samples are balanced, so we do not expect a big improvement in performance of MWT over the moderated $t_m$ test. This is what we see in the operating characteristic (OC) curves in Figure 8b, where MWT and $t_m$ perform similarly, and both are better than the standard $t$ and Welch tests.

4.5 Real data analyses

In comparing the performance of the various procedures with real data, note that the standard $t$ and moderated $t_m$ tend to be biased downwards, so the apparent estimates are not directly comparable. These estimates are given only to indicate that the FDR from MWT can be computed routinely. Figure 9 shows the FDRs from the datasets in Figure 1. The apparent estimates are comparable, except in the ER example, where we expect the standard and moderated statistics to be most biased.

4.6 Simulation from real data

To show the bias problem in real data, we simulate realistic data from the ER status data as follows:

- subtract the group-wise means for each gene;
produce substantial negative bias. For example, at around 1000 the proportion of non-DE genes to unbiased, while the standard true FDRs are comparable, and again (b) shows that MWT is the real (and unknown) unequal variance patterns as well as the dependence structure between the genes. Figure 10a shows the squared error under equal and under unequal variance. We then get a mixture of two groups with the effect size. It is worth noting that in this simulation setup, we preserve the unequal variances. For the example, we simulate two groups with $n_1 = 40$ from the ER positive group and $n_2 = 10$ from the ER negative group ($n_1/n_2$ ratio as observed). We set the proportion of non-DE genes to $\pi_0 = 0.9$, and the effect size to $D = 1$.

It is worth noting that in this simulation setup, we preserve the real (and unknown) unequal variance patterns as well as the dependence structure between the genes. Figure 10a shows the true FDRs are comparable, and again (b) shows that MWT is unbiased, while the standard $t$ and moderated $t_m$ statistics produce substantial negative bias. For example, at around 1000 genes, the bias is about 50% of the target value.

5 DISCUSSION

The key ideas in this article are that (i) unequal group variance is a common problem in practice, (ii) this leads to biased FDR estimates based on standard procedures, (iii) the standard statistical test that deals with unequal variance—the Welch test—lacks power in small samples and (iv) there is a moderated form of the Welch test (MWT) that overcomes these problems. Since MWT performs similarly to the moderated $t_m$ when the group variances are the same, MWT can be used more reliably over a wider set of conditions.

If we use the standard $t$ or $t_m$ tests when the group variances are unequal, the assumed null distribution will have incorrect variance. We then get a mixture of $t$-variates, which have a heavier tail than the assumed $t$ distribution. This typically leads to estimates of FDR which are too optimistic and, consequently, to more false positives than declared. The standard Welch test gets an approximately correct null distribution, but lacks sensitivity in small samples, because of poor estimation of the variance parameters.

Extension of MWT to testing several groups is straightforward: first compute the FDR for testing equality of several variances, then apply it as a weight to the estimated mean-squared error under equal and unequal variance assumptions. Since the detailed method and results are rather extensive, the full description of this idea is given in a separate paper (presently included in Supplementary Report II). Briefly, the results are similar to two-group comparisons: in small samples, the moderated $F$-statistic using weighted mean-squared errors performs better than the standard tests.

Our results show that when the proportion of genes with unequal variance is small or when the sample sizes are balanced, the standard procedures have only small bias; otherwise the bias can be quite large. From our various scenarios, we believe the standard Welch test should not be used in small samples, as it is worse than MWT. The gain in power by MWT over the Welch test is partly due to the weighted formula, which will favor the common variance for a large fraction of genes which have similar variances, and partly due to the pooling of information of standard errors from across the genes.

In conclusion, it might be a good practice to run tests of equal variance as in Figure 1 to check the common variance assumption. We will then know whether the standard tests are reliable. However, in any case the MWT can be expected to work well, so we would recommend its general use in place of the Welch test.

Conflict of Interest: none declared.

REFERENCES


