STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2281–2298
Published online 11 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2109

Significance testing for small microarray experiments

Charles Kooperberg*,†, Aaron Aragaki, Andrew D. Strand and James M. Olson

Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109, U.S.A.

SUMMARY

Which significance test is carried out when the number of repeats is small in microarray experiments can dramatically influence the results. When in two sample comparisons both conditions have fewer than, say, five repeats traditional test statistics require extreme results, before a gene is considered statistically significant differentially expressed after a multiple comparisons correction. In the literature many approaches to circumvent this problem have been proposed. Some of these proposals use (empirical) Bayes arguments to moderate the variance estimates for individual genes. Other proposals try to stabilize these variance estimate by combining groups of genes or similar experiments. In this paper we compare several of these approaches, both on data sets where both experimental conditions are the same, and thus few statistically significant differentially expressed genes should be identified, and on experiments where both conditions do differ. This allows us to identify which approaches are most powerful without identifying many false positives. We conclude that after balancing the numbers of false positives and true positives an empirical Bayes approach and an approach which combines experiments perform best. Standard *t*-tests are inferior and offer almost no power when the sample size is small. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: false positives; type I error; empirical Bayes

1. INTRODUCTION

Research laboratories often perform microarray experiments with only a few (say less than five) repeats. Reasons for the small number of repeats include availability of specimens and economics. While the number of repeats in each experiment is small, commonly the same lab will carry out related experiments: e.g. cell-line A is hybridized three times with and three times without treatment α, the related cell-line B is hybridized four times with and two times without treatment β, and so on. Typically the goal of each individual comparison is to identify

*Correspondence to: Charles Kooperberg, Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, P.O. Box 19024/M3-A410, Seattle, WA 98109, U.S.A.
†E-mail: clk@fhcrc.org

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: CA 74841, CA 53996, NS 42157

Received May 2004
Accepted September 2004

genes that are 'differentially expressed'. Ambitious analyses may want to identify classifiers for the different classes or fit more complicated models.

The limited number of repeats, together with the large variability that even the best microarray platforms have, make small sample comparisons unattractive. A standard t-test for a three-versus-three comparison only has four degrees of freedom. The resulting two-sided test, with $\alpha = 0.05$ and a Bonferoni correction for 10 000 genes requires a t-statistic of 33 or more for significance. The lack of degrees of freedom is really what drives the extremely large significance threshold for t-statistics: the same α and Bonferoni correction for 20 degrees of freedom requires a t-statistic of 6.2 or more while a normal distribution only requires a Z-statistic of 4.6 or more.

To overcome this lack of degrees of freedom we need to combine data one way or another. There are two obvious choices: to get a better estimate of the residual variance for each gene we can combine different genes in the same experiment or we can combine experiments. When genes are combined we can either choose to combine those genes for which the general expression level is similar, or we can choose to combine all genes. If there are closely related experiments, e.g. results on the same tissue under slightly different conditions, we can probably combine experiments. However, just like for combining genes, it may not be clear what we can combine: other tissues? experiments that were carried out earlier? experiments from other labs?

An alternative approach to obtain more power with small experiments is to add a stabilizing constant to the estimate of the variance for each gene or to use some (Bayesian) model for the expression levels. SAM [1] is a methodology that adds a constant to the estimate the variance. The approaches by Baldi and Long [2], Lönnstedt and Speed [3], Smyth [4], and Cui *et al.* [5] are four (empirical) Bayesian approaches. Wright and Simon [6] discuss a closely related frequentist approach.

It is good to realize that permutation tests do not provide an easy way out. For a permutation test with k cases and l controls, without a combination of genes, there are only $\binom{k+l}{k}$ possible levels of the P-value, i.e. 20 levels for the 3-versus-3 and 15 levels for the 4-versus-2 hypothetical experiments mentioned in the first paragraph. For the analysis of a single gene, P-values smaller than 0.05 (even before a multiple comparisons correction) are thus impossible. Some proposed permutation procedures (e.g. Reference [1]) combine the permutation test statistics for all genes. For those procedures we can obtain (somewhat) smaller P-values. While in this paper we focus on P-values obtained from parametric t and normal distributions, we briefly discuss these type of permutation procedures in Section 2.1 and include a comparison in Section 4.4.

In this paper, we do not control for multiple comparisons. In practice, when one carries out tests for many thousands of genes simultaneously, a multiple comparisons correction or a correction of the false discovery (FDR) rate is essential. See Reference [7] for an extensive overview of multiple comparisons corrections. While several of these proposals use permutation arguments to correct for multiple comparisons, permutation typically either requires a substantial number of replicates (that are not available in small experiments), or they require implicit assumptions about genes behaving exchangeable. In either scenario, we believe that only well-calibrated marginal P-values are going to yield good multiple comparison corrected P-values.

P-values have the advantage that there are well-established measures such as type I error and power that can be used to judge the performance of a test. The FDR [8] does not have

such a simple measure, to check whether estimates of the FDR are accurate on a single sample. However, just like for multiple comparison procedures, there are procedures to approximate the FDR from P-values. Thus we hypothesize that for these procedures accurate P-values are a sufficient condition to get an accurate FDR.

2. METHODS

In this paper we compare several methods to analyse two sample comparisons when several parallel experiments are available. We focus here on methods that provide P-values for individual genes from a known reference distribution. (In Section 2.1, we discuss how permutation can be used to obtain P-values for some of these methods.) Assume that we have properly normalized gene expression data x_{ijkl}, where $i=1,\ldots,n$ indicates the gene, $j=1,\ldots,J$ indicates the experiment, $k=1,2$ are the two experimental conditions (that may be different between experiments), and $l=1,\ldots,L_{jk}$ are the number of replicates for condition k in experiment j. We are interested in situations where $L_{jk}<5$ for all j and k. Let μ_{ijk} be the 'true' mean expression level of gene i in experiment j under condition k. Set $\hat{\mu}_{ijk}=\sum_l x_{ijkl}/L_{jk}$ and $s^2_{ijk}=\sum_l (x_{ijkl}-\hat{\mu}_{ijk})^2$.

All of the test statistics that we consider can be written in the form

$$\frac{\hat{\mu}_{ij1}-\hat{\mu}_{ij2}}{\tilde{\sigma}_{ij}}$$

where $\tilde{\sigma}_{ij}$ is some estimate for the variance of $\hat{\mu}_{ij1}-\hat{\mu}_{ij2}$ under the null hypothesis of no differential expression. The traditional test statistics estimate $\tilde{\sigma}_{ij}$ using only the data on gene i and experiment j. The approaches that inflate the variance and those that combine genes also use data on genes i^*, $i^*\neq i$; implicitly to estimate hyperparameters for the empirical Bayes approach that inflates the variance, or explicitly to smooth the estimates for $\tilde{\sigma}_{ij}$ for the estimates that combine genes. Finally the approaches that combine experiments use data on experiments j^*, $j^*\neq j$.

We are comparing the following test statistics.

Traditional test statistics. Several traditional test statistics can be applied to microarray data.

T-statistic. The traditional t-statistic is

$$t_{ij}=\frac{\hat{\mu}_{ij1}-\hat{\mu}_{ij2}}{\hat{\sigma}_{ij}\sqrt{\frac{1}{L_{j1}}+\frac{1}{L_{j2}}}}$$

where $\hat{\sigma}^2_{ij}=(s^2_{ij1}+s^2_{ij2})/(L_{j1}+L_{j2}-2)$, provided $L_{j1}+L_{j2}>2$. The reference distribution is the t-distribution with $L_{j1}+L_{j2}-2$ degrees of freedom, and the main assumption is that for each gene i and experiment j the x_{ijkl} are independent having a normal distribution with variance σ_{ij}, although the t-test is generally considered to be robust against departures from normality.

Welch two sample t-statistic: Welch [9] proposed a two-sample t-statistic when the variances in both groups are different. This statistic has sometimes been used in microarray

analyses. This statistic is defined as

$$w_{ij} = \frac{\hat{\mu}_{ij1} - \hat{\mu}_{ij2}}{\sqrt{\hat{\sigma}^2_{ij1}/L_{j1} + \hat{\sigma}^2_{ij2}/L_{j2}}}$$

where $\hat{\sigma}^2_{ijk} = s^2_{ijk}/(L_{jk} - 1)$. Set

$$c_{ij} = \frac{1}{1 + \frac{L_{j1}\hat{\sigma}^2_{ij1}}{L_{j2}\hat{\sigma}^2_{ij2}}}$$

The reference distribution is approximately a t-distribution with

$$\frac{1}{c^2_{ij}/(L_{j1} - 1) + (1 - c^2_{ij})/(L_{j2} - 1)}$$

degrees of freedom. Note that the number of degrees of freedom for the Welch two-sample test is always smaller than the number of degrees of freedom for the traditional t-statistic.

Inflation of the variance. There exist several methods that inflate the variance, either *ad hoc* (e.g. Reference [1]), using an (empirical) Bayes argument (e.g. References [2–4]), a James–Stein type estimator [5], or a frequentist approach [6]. We include the Limma approach of Smyth [4] as a representative in our simulation study. The reason to use Limma is that (i) it provides explicit P-values (as opposed to, for example, SAM and the approach of Cui *et al.* [10, 5] which require permutation) and (ii) it is easily available as part of the R-Bioconductor project [11]. However, personal communication with authors of some of these approaches suggest that they believe that they work very similar in practice.

Limma: Smyth [4] generalizes the approach from Lönnstedt and Speed [3]. The main assumption in Smyth's model is a prior distribution on the variances σ^2_{ij}:

$$\frac{1}{\sigma^2_{ij}} \sim \frac{1}{d_{0j}s^2_{0j}}\chi^2_{d_{0j}}$$

(We include the index j for the parameters of the prior, as they may be different for different experiments $j = 1, \ldots, J$.) The model also includes priors on the coefficients for each gene in a linear regression model, which in the two sample case reduces to the difference between the mean expression for the two groups. Using methods of moments estimators estimates d_{0j}, s^2_{0j}, and a few other parameters are obtained. An inflated variance $\tilde{\sigma}^2_{ij} = (d_{0j}s^2_{0j} + d_j\hat{\sigma}^2_{ij})/(d_{0j} + d_j)$, where $d_j = L_{j1} + L_{j2} - 2$ is used for a 'moderated t-test' with $d_{0j} + d_j$ degrees of freedom. The approach of Smyth [4] is available from the Bioconductor package Limma. We used Limma with the default options.

Methods combining genes. As for many microarray experiments when the number of replicates L_{jk} is small, the various estimates for the variances are noisy. The effect of this is very few degrees of freedom for the test statistic, which results in reduced power to detect differentially expressed genes. There have been several proposals in the literature to combine the estimates of the variance for several genes to obtain better estimates, so that the resulting test has more degrees of freedom. Typically the assumption that is made is that genes with the same expression level have approximately the same variance. Under this assumption estimates

for the variance can be obtained by smoothing the variance as a function of the expression level. The proposals in the literature differ primarily in how the smoothing is carried out, and whether variances are smoothed jointly or separately for both experimental conditions. In our comparison we include two such proposals.

LPE: Jain *et al*. [12] describe a method they call 'Local Pooled Error test' (LPE). In their approach, let $\hat{\sigma}_{ijk}$ be the initial variance estimate for gene i, experiment j, and condition k, as obtained for the regular t-test. They now proceed by regularizing these estimates for each j and k separately by smoothing the $\hat{\sigma}_{ijk}$ versus $\hat{\mu}_{ijk}$. The assumption being made here is that genes with the same expression level for the same experiment and the same condition have (approximately) the same variance. As the smoothing spline that is used effectively involves averaging a large number of genes, the authors use a normal reference distribution. In our study we have used the implementation by the authors, available in the R-package LPE.

Loess: Huang and Pan [13] make several related proposals. The main difference between their approach and the approach by Jain *et al.* [12] is that they first compute $\hat{\sigma}_{ij}$ and smooth these estimates against $\hat{\mu}_{ij} = \hat{\mu}_{ij1} + \hat{\mu}_{ij2}$. Their simulation results show that, not unexpectedly, for the null-model a normal distribution is appropriate. We reimplemented their approach using a `loess` smoother, and then carrying out a two-sample normal test with equal variances in both groups.

Methods combining experiments. Instead of combining different genes *within* one experiment, we can also combine expression levels of the same gene *between* experiments. This would potentially be useful if we have several smaller experiments, and it is thus reasonable to assume that for each gene the conditional variance for each group in each experiment is approximately the same.

Pooled-t: We define the pooled t-test statistic, combining experiments, as

$$c_{ij} = \frac{\hat{\mu}_{ij1} - \hat{\mu}_{ij2}}{\hat{\sigma}_i \sqrt{\frac{1}{L_{j1}} + \frac{1}{L_{j2}}}}$$

where $\hat{\sigma}_i^2 = \sum_j (s_{ij1}^2 + s_{ij2}^2)/L$ and $L = \sum_j (L_{j1} + L_{j2} - 2)$, provided $L > 0$. The reference distribution is the t-distribution with L degrees of freedom, and the main assumption is that the x_{ijkl} are independent having a normal distribution with mean μ_{ijk} and variance σ_i. Note that, in theory, we would even be able to carry out tests when for a particular experiment j the sample sizes L_{j1} and L_{j2} are 1, provided that $L > 0$ for all experiments combined.

Pooled-loess: We can combine the approach to combine experiments and the approach to combine genes. The possible advantage of such an approach is that we can obtain a more stable estimate of σ combining far fewer genes than what is needed for LPE and the loess approach.

2.1. Permutation P-values

Permutation of the arrays in an experiment can be an alternative to using a parametric reference distribution for a test statistic. Assume that we have a single L_1-versus-L_2 experiment ($J = 1$), and that the test statistic for the ith gene is T_i. To compute the significance of T_i we also compute the test statistics for all genes for each of the $m = 1, \ldots, \binom{L_1+L_2}{L_1}$ permutations of the $L_1 + L_2$ arrays. (One of these permutations will be the original design.) Let T_{im} be the

test statistic for the ith gene for the mth permutation. We can use

$$\sum_{i^*=1}^{n} \sum_{m=1}^{\binom{L_1+L_2}{L_1}} I(T_i < T_{i^*m}) \Big/ \binom{L_1+L_2}{L_1}$$

as an estimate of the P-value corresponding to T_i.

These estimates will be unbiased if (i) each T_i has the same distribution under the null-hypothesis, and (ii) no genes are differentially expressed. The first assumption is not as severe as it appears. When a parametric distribution is used the stronger assumption, that the distributions of each T_i under the null-hypothesis are the same as a particular parametric distribution, is made. The second assumption is much more severe, and it will likely lead to conservative P-values when in fact there are a substantial number of differentially expressed genes [14].

For two of the test statistics which we propose (*Welch* and *Limma*) we know that the first assumption is false, as the number of degrees of freedom differs for each test statistic (*Welch*) or for each different rearrangement of experiments (*Limma*). For the *Pooled-t* and *Pooled-loess* we would have to permute each experiment separately. We will compare this procedure to obtain P-values for the three other approaches to testing that we consider: *t-test*, *LPE*, and *Loess*. Note that for two approaches which we briefly mentioned but not included in our experiments [1, 5] there is no explicit reference distribution, and permutation is required to obtain P-values.

3. EXPERIMENTAL DATA

As experimental data we use affymetrix Mu 11K-A microarrays generated for a series of experiments on Huntington's Disease mouse models. The results of these experiments were reported as a series of related papers [15–17]. For each of these experiments a particular type of mouse with a form of the Huntington's gene inserted was compared to the same type of mouse without the gene mutation at a particular age. At that time mice were sacrificed and gene expression in certain regions of brain tissue were analysed. All comparisons reported in References [15–17] showed some differentially expressed genes, although the amount of differentiation differed considerably between the experiments. For each of the experiments both groups had between 2 and 5 mice. Thus, all our repeats use different samples (sometimes referred to as 'biological repeats') and are not repeat arrays using the same samples (sometimes refereed to as 'technical repeats'), that could be expected to vary less. There are 6595 probe sets (genes) on each array.

For the current paper we reorganize the experiments into groups where we *know* that there are no systematic differences between both experimental groups and some experiments where we know that both experimental groups differ. The experimental line up is shown in Table I. Thus, the experiments Sc1, Sc2, Sc3, Sc4, Ss1, Ss2, and Ss3 are intended to establish that the tests have the right size type I error, and the experiments Dc1, Dc2, Dc3, Ds1, Ds2, and Ds3 are intended to establish the power of the tests.

In our experiment we will analyse these experiments using the analysis methods described in Section 2: for the experiments for which $L_{j1} > 1$ and $L_{j2} > 1$ we will compare seven approaches, for the methods for which $L_{j1} = L_{j2} = 1$ we will compare two methods as well as the 'official' affymetrix P-values. For the methods which combine experiments we compare the more similar

Table I. Organization of the data for our analysis.

Exp.	Tissue	Mouse	Group 1	Group 2	L_{j1}	L_{j2}	Different
Sc1	Cerebellum	DRPLA 26Q	HD	HD	2	2	No
Sc2	Cerebellum	DRPLA 26Q	WT	WT	2	2	No
Sc3	Cerebellum	YAC	HD	HD	3	2	No
Sc4	Cerebellum	YAC	WT	WT	3	2	No
Dc1	Cerebellum	DRPLA 65Q	HD	WT	4	4	Yes
Dc2	Cerebellum	R6/2 12 weeks	HD	WT	2	2	Yes
Dc3	Cerebellum	N171	HD	WT	4	4	Yes
Ss1	Striatum	R6/2 4 weeks	HD	WT	2	2	No
Ss2	Striatum	R6/2 2 weeks	HD	HD	1	1	No
Ss3	Striatum	R6/2 2 weeks	WT	WT	1	1	No
Ds1	Striatum	R6/2 12 weeks	HD	WT	2	2	Yes
Ds2	Striatum	R6/2 6 weeks	HD	WT	1	1	Yes
Ds3	Striatum	R6/2 6 weeks	HD	WT	1	1	Yes

HD: Huntington's Disease mouse, WT: wildtype mouse. Experiments whose code start with a D are expected to have differences between both groups, while those starting with an S are repeats; the second letter of the code refers to the tissue: c stands for cerebellum and s stands for striatum.

experiments within the same tissue. In addition, for the striatum experiments we also look at pooled variances that are obtained by combining the striatum and cerebellum experiments. As the cerebellum experiments have many more degrees of freedom than the striatum experiments (24 versus 4) combining the striatum experiments with the cerebellum experiments had no effect on the cerebellum experiments (results not shown). For all methods we analysed gene expressions that were normalized by two popular methods: the logarithm of the MAS5 average difference summary and the RMA algorithm of Irizarry *et al.* [18]. As all our results were effectively the same for both normalizations, we only report the results on the RMA normalized data. For RMA we normalized all 54 arrays simultaneously; however when we analysed each of the 13 experiments separately, the results were again essentially the same. This is a testament to the robustness of the RMA normalization method. In general, normalization methods had no effect on the results which we present.

4. RESULTS

We are displaying our results as probability–probability plots on a logit-scale (see Figure 1). That is, for a particular method and a particular array let p_i be the two-sided (sometimes called signed) P-values. That is, if p_i is close to 0 there is evidence of under-expression and if p_i is close to 1 there is evidence of over-expression of group one relative to group two. Let $p_{(i)}$ be the sorted p_i. We plot these $p_{(i)}$ (horizontal) against $(1,\ldots,n)/(n+1)$, where $n=6595$, the number of genes on the arrays. We use two criteria for comparing the methods. First, for experiments where both groups are in fact repeats (Sc1, Sc2, Sc3, Sc4, Ss1, Ss2, and Ss3) we would like these plots to follow the identity line. Curves that flatten out are particularly worrisome, as they suggest significantly differentially expressed genes that are in fact false positives. Curves that are more vertical than the identity line suggest statistics that are too conservative: something that is not a concern when there is in fact no difference,

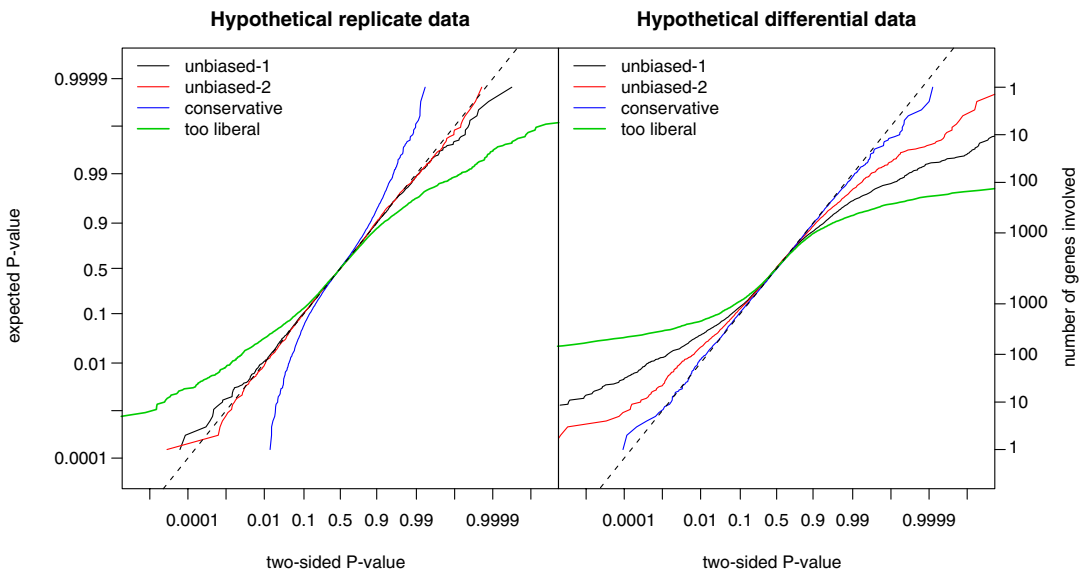

Figure 1. Hypothetical performance of four methods on two data sets. We would prefer the method labelled 'unbiased-1'.

but would likely hurt us when we use the same method to analyse data where some genes are differentially expressed. Second, for experiments where there is a difference between both groups we want the most horizontal curves, among the methods that did not generate a substantial number of false positives for the repeat experiments.

In Figure 1 we show some hypothetical curves. The methods labelled 'unbiased-1' and 'unbiased-2' give good results for the hypothetical replicate data, and among these two 'unbiased-1' performs better on the hypothetical differential data. The 'conservative' method performs acceptable on the hypothetical randomized data, but the conservatism causes the method to call far fewer genes differentially expressed than the two unbiased methods. The 'too liberal' method gives unacceptable results on the replicate data, and its good looking performance on the differentially expressed data would thus be discounted as they likely include many false positives.

4.1. Bandwidth selection

For three of the approaches that we include in our comparison a bandwidth or smoothness parameter needs to be set. In particular, the *Loess* and *Pooled-loess* approaches require the choice of a bandwidth (called `span` in the R implementation) and the *LPE* approach requires the choice of the number of degrees of freedom and a binning parameter. All three of these methods turn out to be extremely insensitive for the choice of these parameters. In Figure 2 we show a comparison of the *P*-value graphs for one differential and one replicate experiment for the *Loess* approach with a wide range of spans. As can be seen, the graphs for different spans are indistinguishable. This is our experience for all three methods that require a bandwidth to be chosen, and for all experiments. Probably this insensitivity should be no surprise, as even

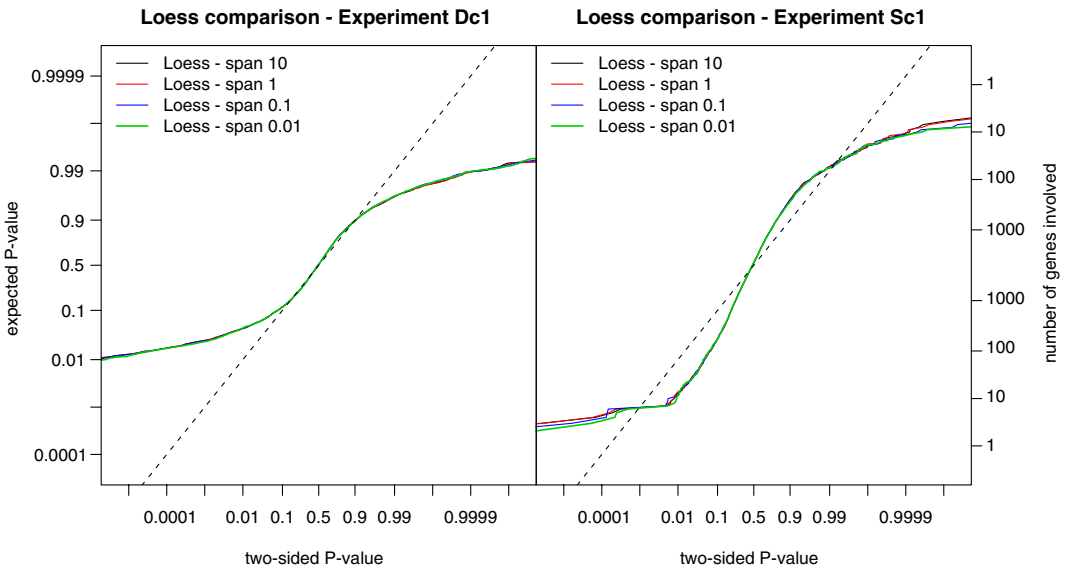

Figure 2. Dependence of the *Loess* approach on the span for two of the data sets.

for the smallest bandwidth these methods still average over a substantial number of genes. As the choice of this parameter is irrelevant, we use a span of 0.1 for the *Loess* and *Pooled-loess* approaches and the default (10 degrees of freedom and 100 bins) for *LPE*.

4.2. Small experiments

In Figure 3 we show the results for four of the five experiments with more than one array for the experiments where both groups are in fact repeats of each other (we omitted experiment Sc4 in this figure to save space). As noted before, these experiments are intended to establish whether these tests all have the right type 1 error. We note from these figures that the three approaches that average variances over genes, *Loess*, *Pooled-Loess*, and *LPE*, are all too liberal. All three methods give approximately 10 times more significant results than appropriate for experiments Sc1, Sc2, and Ss1. For experiments Sc3 and Sc4 (not shown) the *Loess*, and *Pooled-Loess*, method give even worse results, while for these two experiments the *LPE* method performs more reasonably. All other methods give for these experiments either fairly unbiased or sometimes slightly conservative results. The only exception is the *Pooled-t* approach for experiment Sc3, where this approach indicates slightly too many genes that are over-expressed in group one versus group two. We further examined this experiment, and for the genes that are indicated by the *Pooled-t* approach for Sc3 there appears to be a fair amount of difference between both groups. Most of these genes are 'almost' significant for other approaches, and we believe that the variation for the *Pooled-t* approach for Sc3 is random variation.

In Figure 4 we show the results for the four experiments with more than one array for the experiments where both groups are different. We note that the three approaches that

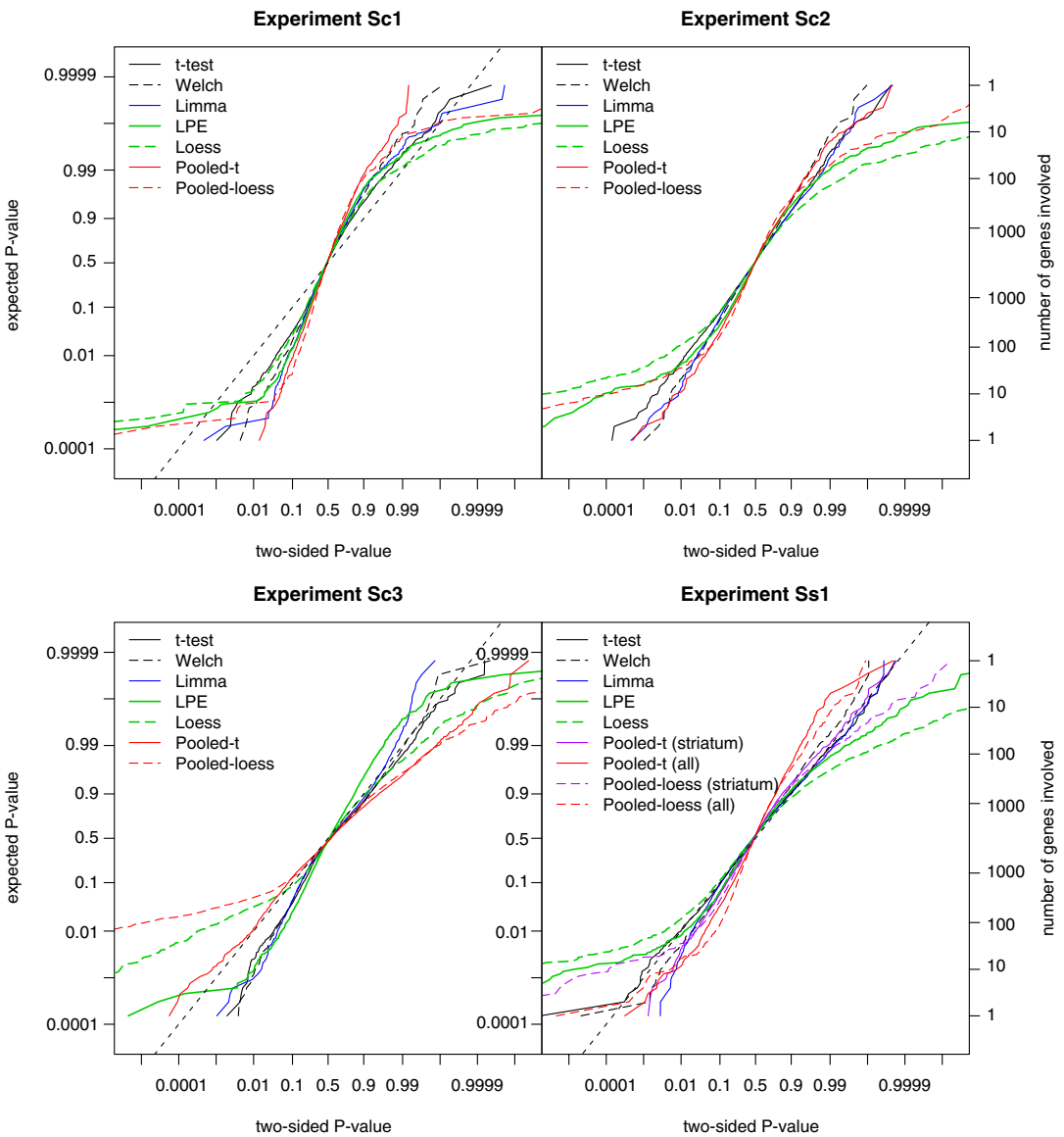

Figure 3. Performance of the various approaches for experiments Sc1, Sc2, Sc3, and SS1. These experiments compare repeat arrays, and there should be no significant number of differentially expressed genes.

yielded biased estimated when the groups are repeats, *Loess*, *Pooled-Loess*, and *LPE*, yield the largest number of positives. However, we know from the repeat experiments that many of these will be false positives. Among the other four methods, the *t-test* and *Welch* approaches, both classical test-statistics, appear to have no power to detect differential expression. The

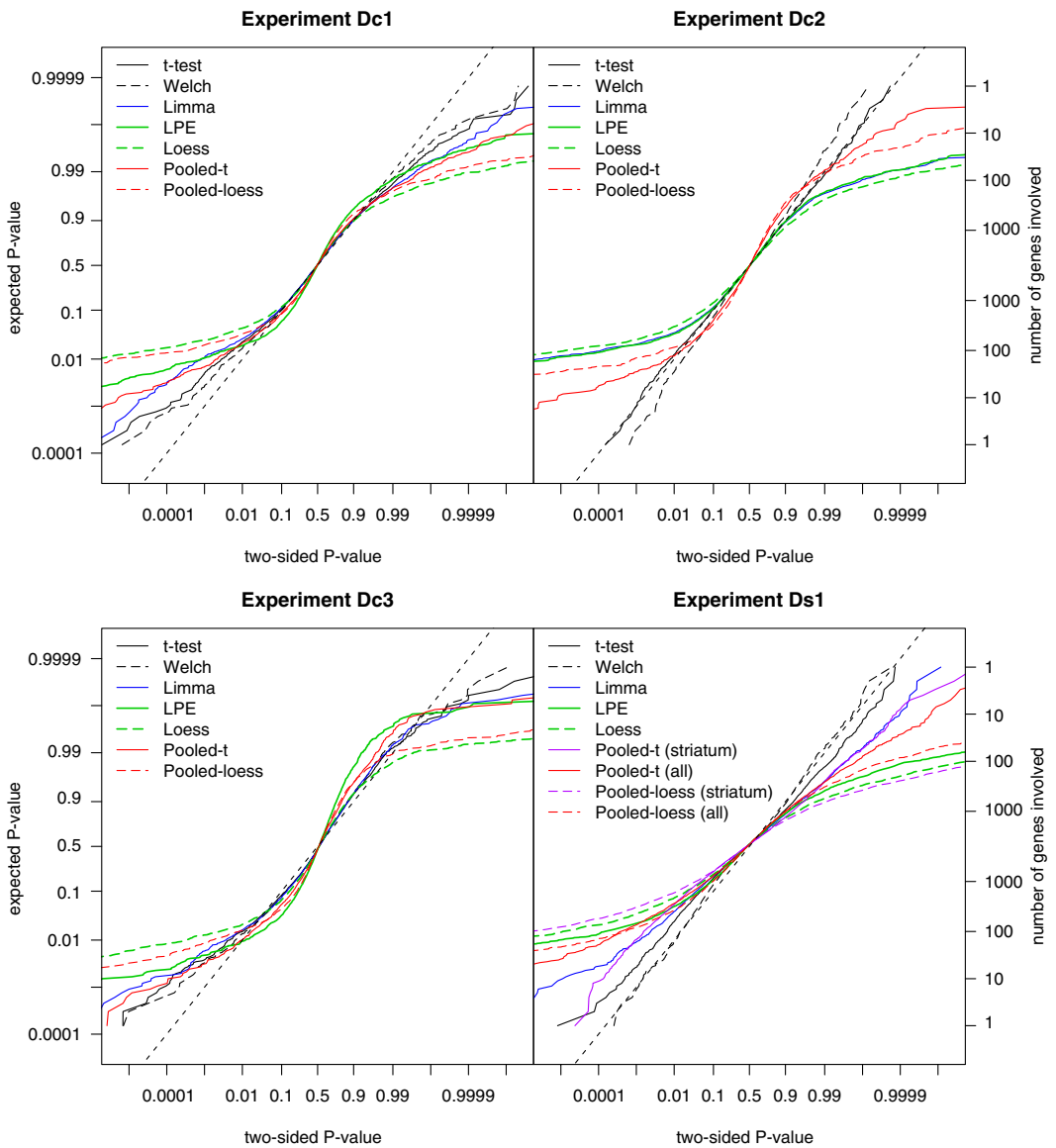

Figure 4. Performance of the various approaches for experiments Dc1, Dc2, Dc3, and Ds1. These experiments compare different mouse lines; thus we prefer methods that indicate many differentially expressed genes.

Pooled-t and *Limma* approaches do better. In fact over these experiments *Pooled-t* slightly outperforms *Limma*. For the *Pooled-t* approach for the striatum experiments with differential expression (Ds1) the power seems to be improved when all arrays, including those that involve

cerebellum, are used over those that only pool the striatum experiments, without biasing the results for the repeat experiment (Ss1).

4.3. *One-versus-one experiments*

Traditionally we would not consider carrying out tests when there are no repeat observations. There is no redundancy in the measurements, and as such, there is no robustness in the experiment: we could not determine whether a single measurement is an outlier, and we have no possibility to check whether any of the underlying assumptions hold. After all, we have no degrees of freedom to estimate the residual variance σ^2. However, using the two approaches that combine different experiments technically allows us to compare two experimental groups without repeats, as long as we can borrow degrees of freedom from other experiments. We compare these methods to the Affymetrix MAS5 estimate of the probability of differential expression, which is provided for single array comparisons.

In Figures 5 and 6 we display the results for the experiments without repeats. We note roughly the same pattern as for the slightly larger experiments: *Pooled-Loess* yields biased results, while the *Pooled-t* yields fairly unbiased results. The power of the *Pooled-t* method is slightly better when all arrays (cerebellum and striatum) are used then when only the striatum arrays are used. The power is very low though, confirming that one-versus-one experiments are not a good idea. The Affymetrix *P*-values do not appear to behave like true *P*-values, and they suggest very many differentially expressed genes, even when the arrays are repeats. These *P*-values are the worst among those studied.

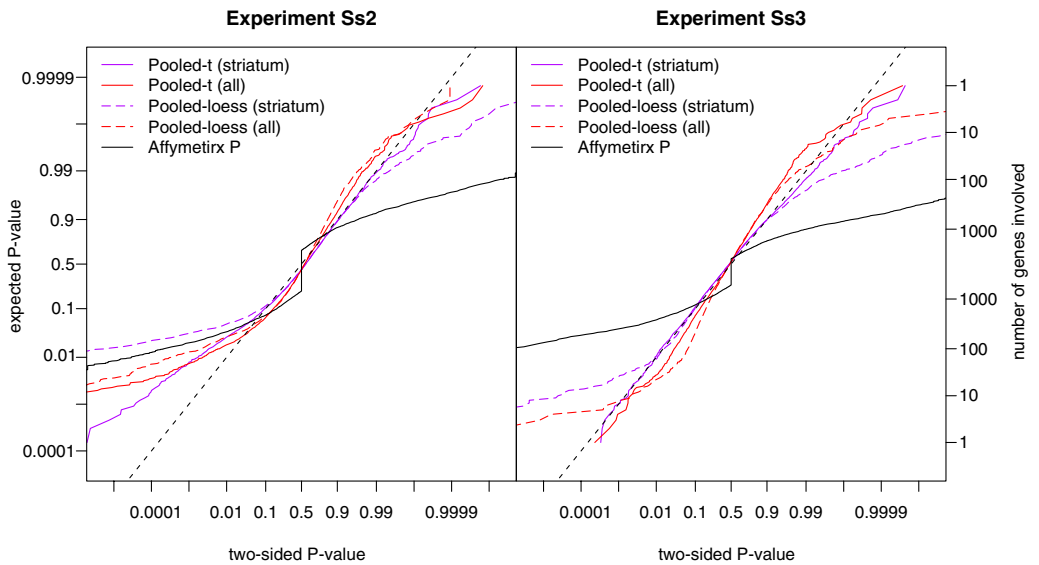

Figure 5. Performance of the various approaches for the experiments Ss2 and Ss3 with one array each. This experiment compares repeat arrays, and there should be no significant number of differentially expressed genes.

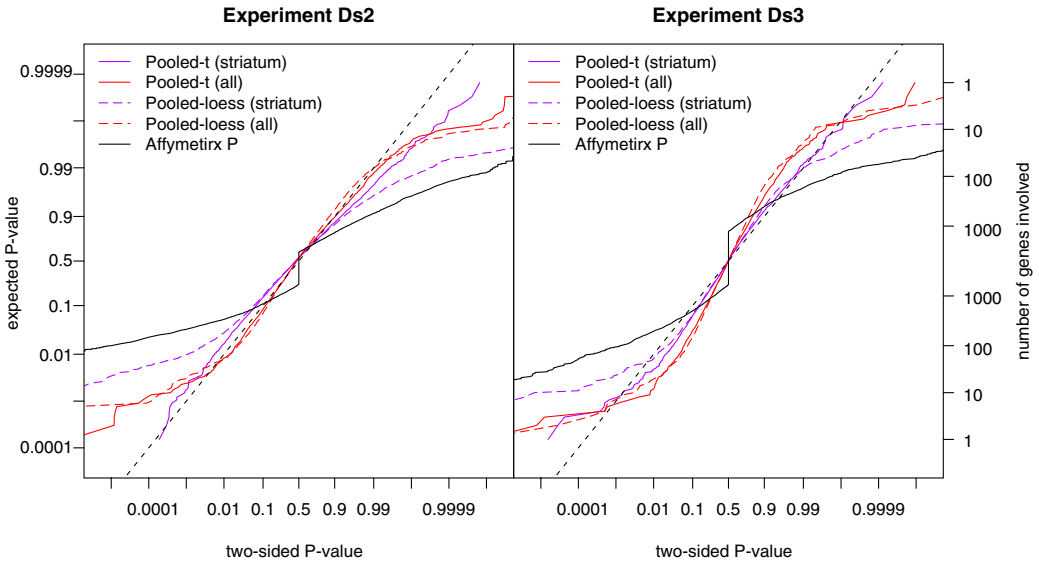

Figure 6. Performance of the various approaches for the experiments Ds2 and Ds3 with one array each. These experiments compare different mouse lines; thus we prefer methods that indicate many differentially expressed genes.

4.4. *Permutation P-values*

As detailed in Section 2.1, an alternative approach to obtaining *P*-values is a permutation approach in which the test statistics for all genes are combined. In Figures 7 and 8 we compare this approach for three of the methods with the (parametric) *Pooled-t* approach. The results for the four experiments displayed in these two figures are similar to those for the other experiments. We notice that the permutation approach for computing *P*-values yields approximately unbiased results, but, as expected, the permutation approach reduces power: the *Pooled-t* approach is consistently more powerful than any of the approaches using permutation.

4.5. *Summary of the results*

In Table II we summarize the results on the mouse data. For each of the seven approaches to computing test statistics compared in Section 4.2 and the three permutation approaches compared in Section 4.4 we counted what fraction of the genes would be significant at (two-sided) significance levels of $\alpha = 1$ and 0.01 per cent. We averaged over the experiments with at least two repeats in each group, separately for the experiments where the two groups are in fact repeats (Sc1, Sc2, Sc3, Sc4, and Ss1) and the experiments where the two groups are different (Dc1, Dc2, Dc3, and Ds1).

Table II confirms our earlier analysis: the *Loess* and *Pooled-loess* approach call very many false positives; *LPE* calls too many false positives at $\alpha = 0.01$ per cent. The remaining methods maintain correct type I error rates. Among these the *Pooled-t* and especially *Limma*

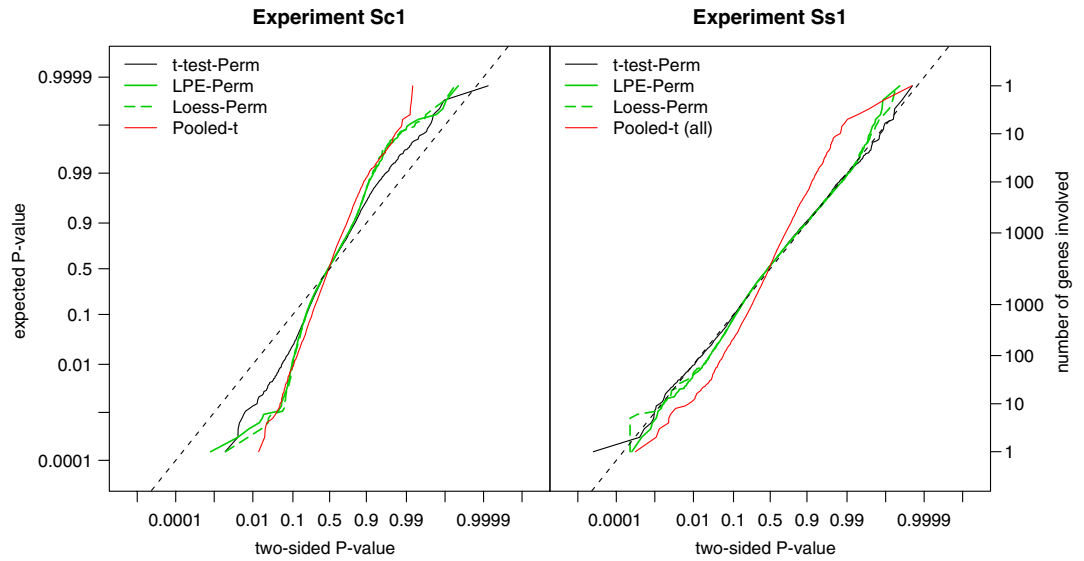

Figure 7. Performance of the permutation approaches and *Pooled-t* for the experiments Sc1 and Ss1. This experiment compares repeat arrays, and there should be no significant number of differentially expressed genes.

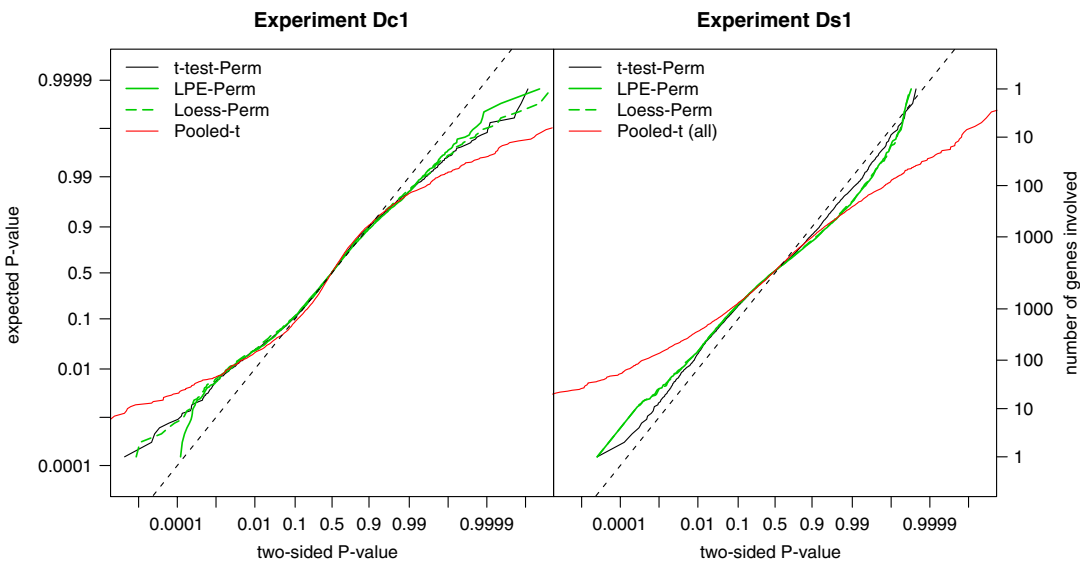

Figure 8. Performance of the permutation approaches and *Pooled-t* for the experiments Dc1 and Ds1. These experiments compare different mouse lines; thus we prefer methods that indicate many differentially expressed genes.

Table II. Percentage of genes that are called significant at different nominal α levels for the mouse data.

	Experimental groups are in fact repeats		Experimental groups are different	
Nominal α	1 per cent	0.01 per cent	1 per cent	0.01 per cent
t-test	0.66	0.006	1.90	0.061
Pooled-*t*	1.06	0.018	2.80	0.254
Welch	0.31	0.006	1.03	0.042
Loess	2.70	0.576	7.29	2.866
Pooled-loess	2.75	0.886	6.35	2.415
LPE	0.73	0.136	4.32	1.467
Limma	0.29	0.006	3.72	0.754
t-test permuted	0.69	0.003	1.95	0.068
Loess permuted	0.46	0.000	2.33	0.027
LPE permuted	0.41	0.003	2.41	0.064

approaches are the most powerful. In particular the permutation approaches do correct the type I error rates, but reduce the power considerably.

5. SIMULATION

To confirm that our results were not dependent on our data set, we conducted a simulation study consisting of four experiments: a four-versus-four experiment and a two-versus-two experiment where there is no differential expression, and a four-versus-four experiment and a two-versus-two experiment where some genes are differentially expressed. All experiments had 10 000 genes. Thus, $L_{i1k} = L_{i3k} = 4$ and $L_{i2k} = L_{i4k} = 2$, for $i = 1, \ldots, 10\,000$, and $k = 1, 2$.

The data was generated as follows. Let $x_{ij1l_1} = \mu_i + Z_{ij1l_1}$ and $x_{ij2l_2} = \mu_i + \delta_{ij2} + Z_{ij2l_2}$ for $i = 1, \ldots, 10\,000$, $j = 1, \ldots, 4$, $l_1 = 1, \ldots L_{ij1}$, and $l_1 = 1, \ldots L_{ij2}$. and $\mu_i \sim \text{Unif}\ [0, 10]$ for all i. The differential expression $\delta_{ij2} = 0$ if $j = 1, 2$ or $i = 1, \ldots, 6000$ and $\delta_{ij2} = \frac{1}{5}(2B_{ij} - 1)G_{ij}$ otherwise, where $B_{ij} \sim \text{Bernoulli}(0.5)$ and $G_{ij} \sim \text{Gamma}(5)$. Thus for experiments 3 and 4, 4000 of the genes have some differential expression, the amount of differential expression varies from very small to substantial.

In the first part of the simulation, we take the variation $Z_{ijkl} \sim \bar{\text{N}}(0, \sigma_i^2)$, where $\sigma_i = (0.3 - 0.02\mu_i)G_i'$ and $G_i' \sim \text{Gamma}(5)$. Thus, the amount of variation depends on the mean μ_i and genes with smaller expression have a larger variance, as is often seen for real gene expression data after a log-transform. However, some of the genes with low μ_i will have a small variance, and some of the genes with a large μ_i will have a large variance.

We generated 10 data sets and averaged the fraction of the genes that was significant at $\alpha = 1$ and 0.01 per cent. The results, shown in Table III, lead to similar conclusions as the mouse data: the *Loess*, *Pooled-loess*, and *LPE* approaches have an increased type I error, and of the other methods the *Pooled-t* and *Limma* approaches are most powerful. Experiment 3 at $\alpha = 0.01$ per cent shows that when there is a substantial percentage of differential expressed genes the permutation approaches can be less powerful than a regular *t*-test.

Table III. Percentage of genes that are called significant at different nominal α levels for the simulated data.

	Tests with no difference				Tests with difference			
	$L_{i1k}=4$		$L_{i2k}=2$		$L_{i3k}=4$		$L_{i4k}=2$	
Nominal α	1 per cent	0.01 per cent	1 per cent	0.01 per cent	1 per cent	0.01 per cent	1 per cent	0.01 per cent
t-test	1.0	0.011	1.0	0.007	9.8	2.114	2.9	0.036
Pooled *t*-test	1.0	0.011	1.0	0.009	11.7	5.239	8.4	2.792
Welch	0.7	0.007	0.3	0.001	8.5	1.282	1.0	0.008
Loess	3.9	0.702	5.0	1.054	14.9	7.532	12.8	5.093
Pooled loess	3.6	0.640	3.5	0.648	14.2	6.806	10.3	3.536
LPE	1.4	0.163	2.4	0.380	8.8	3.267	8.3	2.580
Limma	1.1	0.016	1.0	0.000	11.2	2.487	4.8	0.004
t-test permuted	1.0	0.011	1.0	0.008	9.3	0.345	2.0	0.020
Loess permuted	1.0	0.010	1.1	0.011	8.0	0.339	2.7	0.030
LPE permuted	1.0	0.011	1.1	0.011	5.0	0.244	2.7	0.030

Table IV. Percentage of genes that are called significant by the *Pooled-t* approach at different nominal α levels for the simulated data where the variances differ between experiments.

	Tests with no difference				Tests with difference			
	$L_{i1k}=4$		$L_{i2k}=2$		$L_{i3k}=4$		$L_{i4k}=2$	
Nominal α	1 per cent	0.01 per cent	1 per cent	0.01 per cent	1 per cent	0.01 per cent	1 per cent	0.01 per cent
$a=0,\ r=1.0$	1.0	0.011	1.0	0.009	11.7	5.239	8.4	2.792
$a=0.091,\ r=1.2$	1.1	0.011	1.0	0.015	11.7	5.240	8.4	2.814
$a=0.2,\ r=1.5$	1.1	0.012	1.2	0.022	11.8	5.277	8.5	2.832
$a=0.33,\ r=2.0$	1.2	0.013	1.5	0.031	11.8	5.216	8.7	2.931
$a=0.5,\ r=3.0$	1.4	0.033	2.0	0.088	11.8	5.233	9.1	2.956
$a=0.6,\ r=4.0$	1.6	0.042	2.3	0.136	12.0	5.322	9.4	3.059

When judging Table III we need to keep in mind that in fact for this simulation all four experiments had the same variance for the same gene. Thus the assumptions behind the *Pooled-t* approach are exactly correct. To examine what happens when this assumption is violated we generated additional data for which the set-up is identical as above, except that $Z_{ijkl} \sim \bar{N}(0, \sigma_{ij}^2)$, where $\sigma_{ij} = \sigma_i U_{ij}$, σ_i is as in the first part of the simulation, and $U_{ij} \sim \text{Unif}[1-a, 1+a]$, where a is chosen such to make the maximum possible ratio between σ_{ij} and $\sigma_{ij'}$ equal to r. We again generated 10 data sets and averaged the fraction of the genes that was significant at $\alpha = 1$ and 0.01 per cent. From the results shown in Table IV we note that when $r \leqslant 1.5$ the increase in type I error is small, and that even when $r = 4$ the *Pooled-t* approach still has a smaller type I error than the *LPE* approach.

6. DISCUSSION

Which significance test is carried out when the number of repeats is small in microarray experiments can dramatically influence the results. We set up our experiments so that we could both judge which approaches yield approximately unbiased *P*-values when the experimental conditions are identical, and which approaches are most powerful when both conditions differ. We focused on *P*-values, rather than for example the FDR, as we believe that a 'good' *P*-value will yield a 'good' multiple comparisons correction, and a multiple comparisons adjustment by itself cannot save a procedure that yields too liberal *P*-values.

Our results are striking. As expected, the *t*-statistic and the Welch statistic, arguably the most commonly used statistics, have almost no power when the sample size is small. Alternative methods that combine experiments or genes can have vastly different effects. In our experiments approaches that combine genes using a smoothing of the variance against the expression level yielded large number of false positives. An empirical Bayes approach as well as an approach that pooled related experiments yielded much better results.

Why does combining genes in a smoothing sense lead to false positives? Our explanation is that in fact the variance may depend on the expression level, but that the actual variance for each gene is 'random'. Let us make an assumption similar to that made in the empirical Bayes approach (possibly with a dependence on expression level). If we replace (locally) all variance estimates with an average estimate, as the *LPE* and *Loess* method are effectively doing, the variance will be too large for some genes and too small for other genes. Having some variances too large yields conservatism for those genes, but having too small variances yields false positives for other genes. Thus, the false positives for methods like *LPE* and *Loess* are likely found for genes that have larger variances. The empirical Bayes approaches, such as *Limma* are more subtle, and keep the estimated variances much more like the original variance.

Another conclusion is that borrowing degrees of freedom from other experiments helps. Unfortunately we know of no easy statistical way to judge whether two experiments can be combined within such small experiments. Retrospective analyses of larger experiments may provide insight in which experiments can be combined. We suspect that many experimental conditions are quite robust, and that in fact combining experiments is a much weaker assumption than combining genes. In fact, it is often more the behavior of the gene than of the experiment that determines the variance. It is also useful to realize that most microarray summary measures, such as RMA, are approximately proportional to logarithms of expression levels. Thus, as approximately $\mathrm{var}(\log(Y)) \propto \mathrm{var}(Y)/E(Y)$, the assumption of constant variance on a log-scale is really an assumption of a constant coefficient of variation on the original scale.

We also investigated an approach that extends *Limma* to pool variance estimates between experiments. It turns out that this approach performs almost identical to the *Pooled-t*, as for the pooled data the prior number of degrees of freedom d_o is small compared to the number of degrees of freedom of the experiment.

As we have shown, simple permutation procedures can reduce the bias, but they also reduce power, and these procedures are thus no simple way out. Our experiments were all carried out on Affymetrix arrays. We found out that the actual normalization (RMA or MAS5) had no effect on the results. We hypothesize that most of our results remain valid for two-colour arrays as well.

Our experiments focused exclusively on two-sample comparisons, but as the various tests which we compared primarily differ in their estimates of the residual variance, generalizations to multiple samples via F-tests, or to more complicated modelling situations via linear models, are straightforward for most approaches (LPE and Welch's tests may be the exceptions): all test-statistics which we consider are of the form $(\hat{\mu}_{ij1} - \hat{\mu}_{ij2})/\tilde{\sigma}_{ij}$, which translates directly in $\hat{\beta}/\tilde{\sigma}$ for specific contrasts in multiple sample problems and linear regression. In fact, the Limma approach [4] has already been implemented for those situations.

ACKNOWLEDGEMENTS

We like to thank our HDAG collaborators for allowing us to use their array data and Ziding Feng and Michael LeBlanc for fruitful discussions.

REFERENCES

1. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences U.S.A.* 2001; **98**:5116–5121.
2. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* 2001; **17**:509–519.
3. Lönnstedt I, Speed TP. Replicated microarray data. *Statistica Sinica* 2002; **12**:31–46.
4. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1):3.
5. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. http://www.jax.org/staff/churchill/labsite/pubs/shrinkvariance10.pdf [May 14 2004].
6. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2002; **19**:2448–2455.
7. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003; **18**:71–103.
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
9. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; **29**:350–362.
10. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 2003; **4**(4):210.
11. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5**:299–314.
12. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK. Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* 2003; **19**:1945–1951.
13. Huang X, Pan W. Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Functional and Integrative Genomics* 2002; **2**:126–133.
14. Storey JD, Tibshirani R. Statistical significance for genome wide studies. *Proceedings of the National Academy of Sciences U.S.A.* 2003; **100**:9440–9445.
15. Chan EY, Luthi-Carter R, Strand A, Solana SM, Hanson SA, DeJohn MM, Kooperberg C, Chase KO, Young AB, Leavitt BR, Cha JJ, Aronin N, Hayden MR, Olson JM. Increased huntingtin protein length reduces the number of polyglutamine-induced gene expression changes in mouse models of Huntington's disease. *Human Molecular Genetics* 2002; **11**:1939–1951.
16. Luthi-Carter R, Hanson SA, Strand AD, Bergstrom DA, Chun W, Peters N, Woods AM, Chan EY, Kooperberg C, Young AB, Tapscott SJ, Olson JM. Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain. *Human Molecular Genetics* 2002; **11**:1911–1926.
17. Luthi-Carter R, Strand AD, Hanson SA, Kooperberg C, Schilling G, LaSpada A, Merry D, Young AB, Ross CA, Borchelt DR, Olson JM. Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects. *Human Molecular Genetics* 2002; **11**:1927–1937.
18. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; **4**: 249–264.

