BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 1 2007, pages 38–43
doi:10.1093/bioinformatics/btl548

Gene expression

Estimating *p*-values in small microarray experiments

Hyuna Yang and Gary Churchill*

The Jackson Laboratory, Bar Harbor, ME 04609, USA

Received on May 1, 2006; revised on October 12, 2006; accepted on October 22, 2006
Advance Access publication October 30, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Microarray data typically have small numbers of observations per gene, which can result in low power for statistical tests. Test statistics that borrow information from data across all of the genes can improve power, but these statistics have non-standard distributions, and their significance must be assessed using permutation analysis. When sample sizes are small, the number of distinct permutations can be severely limited, and pooling the permutation-derived test statistics across all genes has been proposed. However, the null distribution of the test statistics under permutation is not the same for equally and differentially expressed genes. This can have a negative impact on both *p*-value estimation and the power of information borrowing statistics.

Results: We investigate permutation based methods for estimating *p*-values. One of methods that uses pooling from a selected subset of the data are shown to have the correct type I error rate and to provide accurate estimates of the false discovery rate (FDR). We provide guidelines to select an appropriate subset. We also demonstrate that information borrowing statistics have substantially increased power compared to the *t*-test in small experiments.

Contact: garyc@jax.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray technology has made possible the simultaneous expression profiling of thousands of genes but cost and other considerations often limit the number of replicated samples in an experiment. Testing for the differential expression of many genes on small numbers of samples is problematic. However it is possible to leverage the multiplicity of genes to our advantage.

The most widely used statistical method for comparing two groups is the *t*-test and, not surprisingly, it is common in microarray data analysis. However, problems with the *t*-test are known to arise when the number of observations per gene is small due to instability in the estimation of gene specific variances (Tusher *et al*., 2001; Smyth *et al*., 2003). Microarray data provide information from thousands of genes, and the combined information can be used to obtain stable variance estimates. Test statistics that utilize across-gene information, such as F_s(Cui *et al*., 2005) and B (Lonnstedt and Speed, 2002; Smyth, 2004), have been developed but the null distributions of these statistics are not known. Permutation analysis is the best available method to obtain *p*-values.

To estimate a *p*-value for gene g ($g = 1, \ldots, G$) using permutation analysis, one first calculates the observed test statistic T_g. Then one redistributes the observations among the test and control groups and recalculates the test statistic, T^*_{g1}. Depending on the size of the experiment, one can either enumerate all possible permutations or generate a random sample of permutations. The *p*-value is estimated by counting the number of $T^*_g = \{T^*_{gi} : i = 1, \ldots, M\}$ that are greater than or equal to T_g and dividing by the total number of permutations, M. When the sample size is 3 per group, there are 20 possible randomizations. There are 10 distinct values of the test statistic ignoring the sign, and thus the smallest possible *p*-value is 0.1. Similarly for an experiment with 5 samples per group, there are 126 distinct values of the test statistic under permutation, and the smallest possible *p*-value is 0.008. Thus small sample sizes can severely restrict the possible *p*-values that can beobtained in a permutation analysis. The need for multiple test adjustments exacerbates the problem.

To overcome this problem, pooling of permutation-derived test statistics across all genes has been proposed (Storey and Tibshirani, 2003). We let $T^* = \cup_{g=1}^{G} T^*_g$ and use entire set of test statistics across all genes as a null distribution to estimate *p*-values for each gene. As noted by Storey and Tibshirani (2003), the null distribution of each differentially expressed gene might be different, and thus the distribution of the pooled sample distribution represents a mixture. Differential expression tends to increase the variance of the null distribution. The pooled null distribution from experiments with many differentially expressed genes will have heavier tails, and the *p*-values estimated from this distribution will tend to be conservative. Storey and Tibshirani argued that the mixture distribution should not be a problem for false discovery rate (FDR) estimation, however our simulation study shows that it can be problematic in small experiments where pooling is essential.

Xie *et al*. (2005) and Fan *et al*. (2005) noted problems with the permutation test and proposed a modification that involves pooling *p*-values over a selected subset of the data. Fan *et al*. (2005) use individual gene tests based on the *t*-distribution to obtain the subset. Xie *et al*. (2005) select a subset by removing the same number of genes as are estimated to be differentially expressed (DE) genes. However estimating the number of DE genes is a challenging problem. We have adopted these ideas and develop them further. We propose several strategies for obtaining subsets of genes for pooling.

To assess the validity of the resulting *p*-values we consider three properties. The first is the type I error rate. For any α in (0,1), we require that the test of a true null hypothesis will yield a *p*-value less than α with probability no larger than α. This is an essential condition to be met. The second property is the accuracy

*To whom correspondence should be addressed.

of the estimated FDR obtained from the *p*-value distribution. This is important due to the widespread use of FDR to correct for multiple testing in microarray experiments. Finally, on the condition that a method yields the right type I error rate, we prefer a method with good power.

In this paper, we study two-condition comparisons but we note that the proposed methods are readily extended to multiple conditions. We demonstrate the performance of the different *p*-value estimation methods using the *t*-statistic and two information borrowing statistics. We define the t_s statistic to be the two-condition comparison form of the F_s statistic, and t_b is the moderated *t*-statistic (Smyth, 2004). These modified forms of the *t*-statistic 'borrow information' from other genes, but they differ in the methods used to estimate error variances.

2 METHODS

2.1 Test statistics for two condition experiments

The *t*-statistic is used to test whether a gene is DE or equally expressed (EE) between two conditions. The estimated variance in the denominator of the *t*-statistic is based only on data from one gene. It can be unstable and may result in poor overall performance of the *t*-test. It would be desirable to pool variance estimates across all genes but simple averaging does not allow for gene specific variances (Cui *et al*., 2005).

Each gene in a microarray experiment can have its own unique variance. This may be a consequence of biological or technical factors but it is clear from our experience that variances are variable across genes to a greater extent than expected due to statistical errors of estimation. To derive stable gene specific variance estimates, we can borrow information across genes by shrinking the variance estimates toward a prior value or toward their bias-corrected geometric mean. When the true variances are highly variable it is desirable to shrink less. When the true variances are similar we should shrink more. In this way the new variance estimates adapt to the degree of heterogeneity of variances.

The statistics t, t_s and t_b differ in how each estimates the variance. Let S_t, S_s and S_b be the variance estimates used to compute the t, t_s and t_b statistics, respectively. S_b and S_t have a simple linear relationship $S_b = a + bS_t$, where $a = d_0 s_0^2/(d_0 + d_g)$ and $b = d_g/(d_0 + d_g)$ when s_0^2 is a prior estimator of variance, and d_0 and d_g are degrees of freedom of s_0^2 and S_t, respectively (Smyth, 2004). S_s is derived as a James–Stein estimator of variance on the log scale, and $\log(S_s)$ and $\log(S_t)$ have a linear relationship (Cui *et al*., 2005). S_b and S_s both are empirical Bayes estimators.

2.2 Estimation of *p*-values

Permutation *p*-values were proposed by Fisher (1935) as a measure of 'strength of evidence' against a simple null hypothesis. For small sample experiments, one can list all possible arrangements of the data into treatment groups and measure the extent to which the observed configuration is extreme. In microarray experiments, permutation of observations from a single gene can yield exact and unbiased *p*-value estimates for that gene. However, since the null distributions from each gene might be different, pooling the permuted data test statistics across genes cannot be guaranteed to yield correct *p*-values. Xie *et al*. (2005) have observed that permutation test statistics can overestimate the tails of the null distribution resulting in conservative inference. The problem arises because the permutation distribution of DE genes will have a larger variance than that of EE genes. Ideally we would derive null distributions individually for each gene and circumvent this problem. However, when the number of samples is small, the resulting *p*-values are too sparse, and the smallest attainable *p*-value can be too big. Thus there is a need to obtain a sufficient number of permuted test statistics to obtain an accurate estimation of the null distribution.

Follow the suggestion of Xie *et al*. (2005) and Fan *et al*. (2005), we consider using a selected subset of the data for the permutation analysis. We will address the subset selection procedure further in section 3.2, and first investigate the *p*-value estimation methods. We propose two strategies.

- Subset selection before permutation: for each gene g calculate the *t*-statistic and remove the gene if the absolute value is bigger than the α-level critical value of the *t*-distribution. Using remaining genes ($j \subset \{1, \ldots, G\}$), conduct a permutation analysis and pool the resulting test statistics to obtain $T_j^* = \{T_{j1}^*, T_{j2}^*, \ldots, T_{jM}^*\}$. Then compute estimated *p*-values for all genes $g = \{1, \ldots, G\}$ based on the set T_j^*.
- Subset selection after permutation: for all genes ($g = \{1, \ldots, G\}$) conduct a permutation analysis to obtain test statistics $T_g^* = \{T_{g1}^*, T_{g2}^*, \ldots, T_{gM}^*\}$, and form a pool using only the T_j^* from genes whose absolute *t*-statistics are bigger than the α-level critical value of the *t*-distribution. Compute estimated *p*-values for all genes from this set.

The test statistic T need not be the *t*-statistic, and in this paper t, t_s and t_b statistics are studied. Both methods use the standard *t*-statistic to define the subset, but they differ in the stage at which we select the subset. This difference is only relevant to the information borrowing statistics. If a test statistic is calculated based on data from each individual gene, as is the case with the standard *t*-test, the two subset selection methods are identical.

One advantage of using the *t*-distribution to define a subset is that the criteria for selecting a subset will be sensitive to the number of true DE genes in the data. Optimal subset selection should depend on the number of DE genes; when there are many (few) DE genes, we should remove many (few) genes from the set to be pooled. Removing too many or too few genes can alter the estimated null distribution as we illustrate below. The choice of an appropriate percentile of the *t*-distribution is investigated in our simulations, and here we use the $\alpha = 0.10$ (two tailed) critical value. Another advantage is that it provides a reasonably robust selection criteria that does not rely on permutation analysis.

2.3 Simulation design

We conducted a simulation to compare the performance of the different test statistics and *p*-value estimation methods. We focus on small sample size experiments having two conditions, test versus control, and sample sizes of 3.5 or 10 per condition. We generated data from 10 000 genes and varied the proportion of DE genes as 0.01, 0.1 or 0.5. Control group data and test group data for EE genes were drawn from a $N(0, \sigma_g^2)$ distribution, where σ_g^2 could be constant, moderately variable or highly variable. For the constant variance case, we set $\sigma_g^2 = 1$, otherwise we sampled random variances for each gene from an inverse Gamma distribution. Note that when $\sigma_g^2 \sim 1/\text{Gamma}(a,a)$, the mean is $E(\sigma_g^2) = a/(a-1)$ and the variance is $\text{Var}(\sigma_g^2) = a^2/((a-1)(a-2))$. For moderately variable variances we used $a = 30$, and for highly variable variances we used $a = 5$. Test group data for DE genes were drawn from a $N(\mu_g, \sigma_g^2)$ distribution, where μ_g were sampled from a Gamma(*a*,*b*) distribution. To allow the variance of $\mu_g (= \frac{a}{b^2})$ to increase with the mean of $\mu_g (= \frac{a}{b})$, and mean of μ_g to be 0.5, 1, 2 and 4, respectively, we sampled $\mu_g \sim \text{Gamma}(4, 8)$, $\mu_g \sim \text{Gamma}(4, 4)$, $\mu_g \sim \text{Gamma}(8, 4)$ and $\mu_g \sim \text{Gamma}(17.5, 3.5)$. Supplementary Figure 1 illustrates the distributions of σ_g^2 and μ_g under each parameter setting. In total we consider 108 different parameter settings using three sample sizes, three proportions of DE genes, three degrees of variance heterogeneity and four average fold changes, in all combinations. We discuss only the most interesting cases below.

3 RESULTS

3.1 Simulation results

We generated data as described in Methods and computed *p*-values using 15 different methods. Five methods were used to estimate *p*-values from *t*-statistics: the *t*-distribution (*tab.t*), permutation of

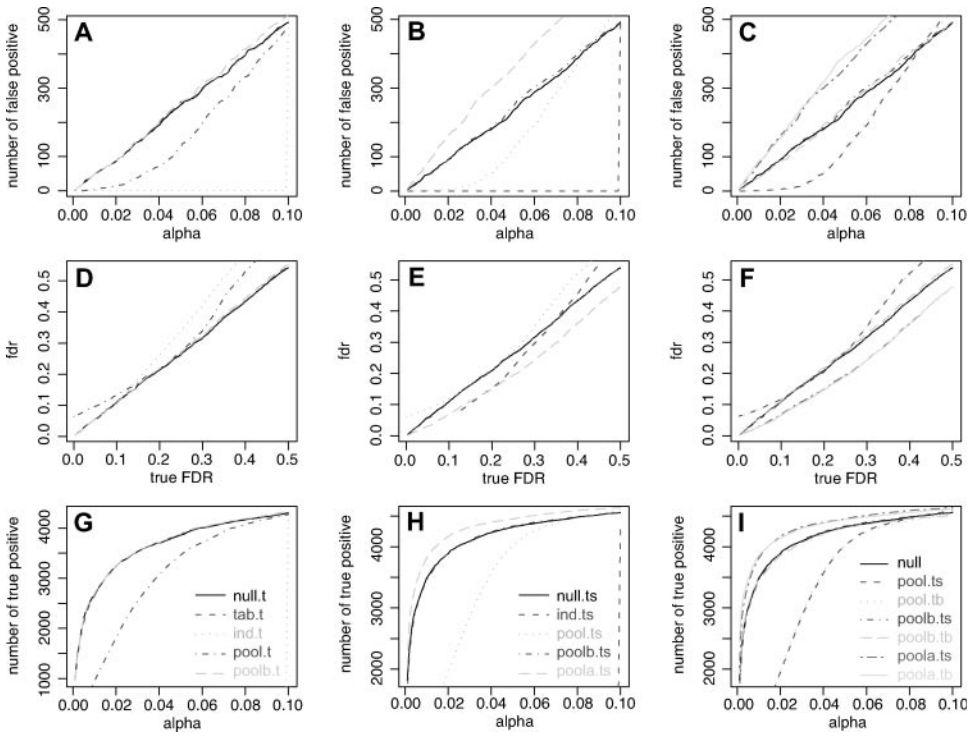

Fig. 1. Comparisons of 15 different *p*-value estimation procedures. *p*-values were obtained for simulated data from an experiment with sample size three per group, 5000 DE genes with mean $\log_2$ fold change of 4, and 5000 EE genes. (**A–C**) show the numbers of false positives. (**D–F**) show true FDR versus estimated FDR using the *qvalue* function and (**G–I**) show the numbers of true positives. (A, D and G) and (B, E and H) compare results using t and t_s statistics, respectively. (C, F and I) compare thet$_s$ and t_b statistics.

individual genes with no pooling (*ind.t*), permutation of EE genes with pooling (*null.t*), permutation of all genes with pooling (*pool.t*) and permutation with a subset of genes selected for pooling (*poolb.t*). As we noted earlier, subset selection before or after permutation are identical for the *t* test, which does not share information across genes. The permutation of EE genes (*null.t*) is only possible in simulations as, in practice, the EE or DE status of a gene is not known. It provides a truth standard in the simulation. We also used five different methods to derive *p*-values from each of the t_s and t_b statistics. We denote the t_s methods as *null.ts*, *ind.ts*, *pool.ts*, *poolb.ts*, *poola.ts* and similarly for t_b. The *t*-distribution is not appropriate for information borrowing statistics and subsetting before (*poolb.ts*) or subsetting after (*poola.ts*) are distinct.

To assess performance, we considered the true type I error rate, the accuracy of FDR estimation and the power of each method. Because the *p*-values *null.t*, *null.ts* or *null.tb* are obtained on the correct null distribution, we used these asa reference.

Type I error rate We compared the number of false positives obtained in each simulation to its expectation which is the number of EE genes times the significance level α. For example, when the number of EE genes is 1000, we expect 10 false positives at the α level .01. If the observed number of false positive results is greater than 10, the estimated *p*-values are liberal and if it is smaller, the *p*-values are conservative. A conservative test may be acceptable, but there is likely to be a corresponding loss in power. Liberal test are regarded as unacceptable.

Figure 1A–C illustrates the case of an experiment with sample size of 3 per group, 5000 DE genes with a mean $\log_2$ fold change of 4, and highly variable variances. We observed that pooled *p*-values (*pool.t*, *pool.ts* and *pool.tb*) result in conservative tests. Selection of genes for pooling after permutation (*poola.ts* and *poola.tb*) results in under-estimation of *p*-values and thus the test is liberal. Selecting the subset before permutation (*poolb.ts* and *poolb.tb*) results in the expected type I error rate. We note that t_s and t_b showed similar performance. Similar results were obtained under the other parameter settings. However, when the number of DE genes and the mean level of differential expression decreased or the experiment size increased, the differences among the *p*-values from different procedures were less apparent.

False discovery rate FDR has become the standard method for establishing significance in the multiple testing context of microarray data. FDR relies on properties of the *p*-value distribution and is estimated under the assumptions that *p*-values of EE genes follow a uniform distribution and those of DE genes are stochastically smaller (Storey, 2002). We examined our *p*-value estimation methods from the perspective of obtaining accurate FDR estimates. Figure 2 shows histograms of *p*-values from four of these methods obtained under the same conditions as the simulations in Figure 1. Histograms of pooled *p*-values (*pool.t* and *pool.ts*) show a slightly U-shaped density with too few moderate (0.25 to 0.75) *p*-values. The *poolb.t* and *poolb.ts* estimates (as well as *null.t* and *null.ts*) are uniform across the entire right half of the histogram ($P > 0.25$). This suggests that FDR estimates obtained from pooled *p*-values

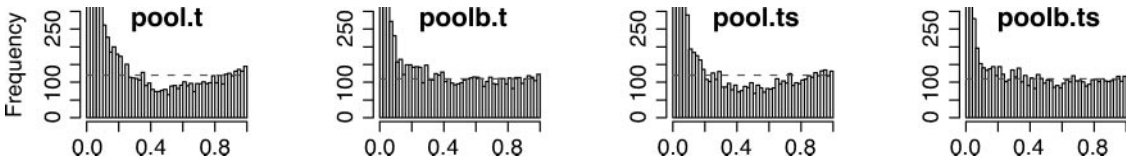

Fig. 2. Histograms of *p*-values obtained using estimation procedures *pool.t*, *poolb.t*, *pool.ts* and *poolb.ts* on simulated data from an experiment with sample size 3 per group, 5000 DE genes with mean $\log_2$ fold change 4, and 5000 EE genes. Histograms are truncated to 300.

may not be as reliable as those obtained with subset selection before permutation. Results using the t_b statistic look identical to those from the t_s statistic.

In order to validate this expectation we calculated the true FDR and the estimated FDR using the *qvalue* function (Storey, 2002) written in R language (Ihaka and Gentleman, 1996). It is desirable that the true and estimated FDR should agree. Figure 1D–F indicates that FDR is overestimated by pooled *p*-values, consistent with the findings of Xie *et al*. (2005). With selection after permutation, FDR is under-estimated. The selection before permutation procedures provide accurate FDR estimates and again t_s and t_b show similar performance.

We also examined the standard deviations of FDR estimates for each parameter setting to assess the precision of the simulation study. We generated 200 independent datasets from a parameter setting, calculated *p*-values, *q*-values and then standard deviations of *q*-value corresponding to the $\alpha = 0.20, 0.10, 0.05, 0.01$ two-tailed critical values of the *t* distribution. The standard deviation of each *q*-value at each percentile was surprisingly small ($<10^{-6}$), confirming the consistency of the previous result.

Power to detect DE Here we consider the true positive rate (power) of each method. Figure 1G–I shows the numbers of true positives. We found that subset selection after permutation yielded the greatest power, but because these tests were liberal we did not consider this to be relevant. The pooled *p*-values are conservative and have low power. For each of the three test statistics, t, t_s and t_b, the number of positive results obtained with subset selection before permutation agrees well with tests based on the true null distribution. The information borrowing statistics t_s and t_b provide the best power.

Based on these simulations, we can conclude that the pooled *p*-value is conservative, that tests based on a subset of genes selected before permutation perform best consistently and that the information borrowing statistics provide the best power.

3.2 Thresholds for subset selection

In the preceding simulations, we compared *p*-value estimation methods and used the $\alpha = 0.10$ critical value of the *t*-distribution to define the selected subsets. We have determined that subset selection before permutation provides the most appropriate *p*-value estimates but did not examine the effect of the criteria for subset selection. To address this question, we reanalyzed the simulated data using the $\alpha = 0.20, 0.15, 0.10, 0.05, 0.01, 0.001$ critical values of the *t*-distribution as threshold values for subset selection. For each parameter setting, we computed *p*-values using different thresholds for subset selection. Table 1 shows the numbers of genes retained in the selected subsets from two simulated data sets each having 5000 DE genes with mean $\log_2$ fold changes of 4 and 0.5, respectively.

Table 1. Numbers of genes remaining after the subset selection from two data sets; each has 5000 DE genes with a mean $\log_2$ fold change of 4 (Data 1) and 0.5 (Data 2). Six critical values, $\alpha = 0.20, 0.15, 0.10, 0.05, 0.01$ and 0.001, were used to select the subsets

α	0.20	0.15	0.10	0.05	0.01	0.001
Data 1	4148	4454	4818	5399	6963	9124
Data 2	7659	8218	8748	9318	9856	9988

More genes are removed from the pool when the data have more DE genes or higher mean $\log_2$ fold change. Thus the subsetting is adaptive to these features of the data. We fit LOWESS curves to the difference between *null.ts* and *poolb.ts* *p*-values (Figure 3A and E). We can see that using $\alpha = 0.001$ or 0.01 critical value of the *t*-distribution as a threshold yields *p*-values that are different from the true null distribution. This is a consequence of failure to remove DE genes. We see that using the $\alpha = 0.20$ critical value of the *t*-distribution as a threshold also yields a conservative result. This can be explained by the behavior of the t_s statistic. Supplementary Figure 2 shows that genes removed tend to have larger variance than those that are not removed. Thus, as we remove more genes, greater homogeneity of variances among the remaining genes leads to a greater shrinkage and to a conservative result. In summary, trimming too many genes or trimming too few both perturb the null distribution and can result in conservative tests.

To identify an optimal threshold, we fit LOWESS (Cleveland, 1979) curves to the t_s statistic versus $-\log$(*p*-value) (Fig. 3B and F). Here we only used *pool.ts* and *null.ts*. Note that $-\log$(*p*-value) from *null.ts* exponentially increases as t_s increases. This is the pattern that we expect when the correct threshold is used. $-\log$(*p*-value) from *pool.ts* also increases as t_s increases, but there is an inflection at $t_s \approx 2$. When $t_s < 2$, the *pool.ts* and *null.ts* are quite similar, but $t_s > 2$ the *pool.ts* *p*-values tend to be bigger than *null.ts* *p*-values. We marked the critical values of the *t*-distribution along the LOWESS curves in Figure 3B and F using vertical lines. The $\alpha = 0.10$ critical values of the *t*-distribution is quite close to the inflection point, suggesting that this is a reasonable threshold for subset selection.

The numbers of true and false positives obtained using *poolb.ts* at each critical value are showing in Figure 3C and G, D and H, respectively. Again we see that *p*-values obtained from threshold below the $\alpha = 0.05$ critical values are too conservative, and that thresholds above the $\alpha = 0.15$ critical value are also slightly conservative. When the mean $\log_2$ fold change of DE genes is 0.5, *p*-values obtained using $\alpha = 0.20$ critical value of the *t*-distribution as a threshold shows the largest deviation from the null. Although the difference is not large compared to the data in which the mean $\log_2$ fold change of DE genes is 4, this indicates that removing too many genes is not desirable.

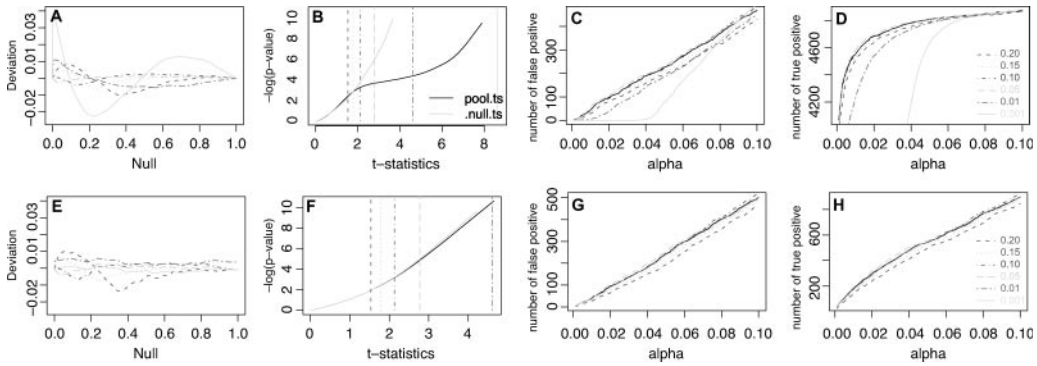

Fig. 3. Comparisons of different thresholds for subset selection: (**A**–**D**) show results from simulated data in which the mean $\log_2$ fold change of DE genes is 4. (**E**–**H**) show result from simulated data in which the mean $\log_2$ fold change of DE genes is 0.5. A and E show LOWESS curve fitting to *null.ts* *p*-value versus the difference between *poolb.ts* and *null.ts* *p*-value. Six different thresholds (α = 0.20,0.15,0.10,0.05,0.01,0.001 critical values of the *t*-distribution) were used to obtain *poolb.ts* *p*-value. B and F show LOWESS curve fitting to the t_s statistic versus $-\log$(*p*-value). *null.ts* and *poolb.ts* were used to obtain *p*-value. (C and G, D and H) show the numbers of false positives and true positives, respectively.

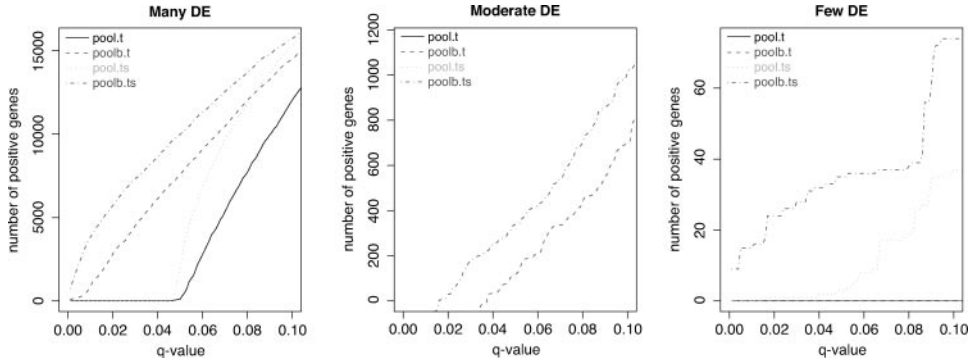

Fig. 4. Numbers of detected genes as a function of the *q*-value from three real microarray data having many, moderate numbers, and few DE genes. Five different *p*-value estimation procedures (*tab.t*, *pool.t*, *pool.ts*, *poolb.t* and *poolb.ts*) were applied to datasets and the $\alpha = 0.10$ critical value of the *t*-distribution was used to create the subset.

We compared different thresholds for other simulation parameters and found that the $\alpha = 0.10$ critical value of the *t*-distribution works well in all cases that we considered in this simulation study. Thus, we recommend $\alpha = 0.10$ critical value of the *t*-distribution as a threshold, however it may be desirable to fit the LOWESS curve to the pooled *p*-value estimates obtained from a given dataset and check for the location of the inflection point.

In summary, simulation studies show that *p*-values based on subset selection before permutation with the $\alpha = 0.10$ critical value of the *t*-distribution as a threshold performs better than other *p*-value estimation methods considered here.

3.3 Applications to Real Data

In this section we consider the behavior of *p*-values obtained with pooling and with the subset selection before permutation procedures using real data. Data from Affymetrix MOE430v2 arrays, run in the gene expression facility at The Jackson Laboratory, are available at http://www.jax.org/staff/churchill/labsite/datasets. We chose three microarray experiments to represent cases with many, moderate numbers, and few DE genes, respectively. In each case we computed five *p*-values, *tab.t*, *pool.t*, *pool.ts*, *poolb.t* and *poolb.ts*, using the $\alpha = 0.10$ critical value of the *t*-distribution as a threshold for the subset selection. Table 1 shows the numbers of genes remaining after subset selection. For data with more DE genes, the selected subset is smaller but in all cases the numbers are more than adequate to obtain precise estimation.

Histograms of *p*-values from each dataset and each of five methods are provided in Supplementary Figure 3. Compared with the pooled *p*-value, the selected subset *p*-values have a sharper peak, a wider uniform area and a higher estimated proportion of EE genes (π_0).

Figure 4 shows the number of genes declared DE as a function of the *q*-value. We can see that the number of detected genes using the pooled *p*-value is quite small for small *q*-values and that it abruptly increases as the *q*-value is raised. The number of detected genes using the subset selection before permutation method increases smoothly. The subset selection before permutation method always yields the greatest number of detected genes compared with other methods. Table 2 shows number of genes declared as DE using *p*-value = 0.001, *q*-value = 0.01 and *q*-value = 0.05 as critical values for detection.

4 DISCUSSION

We have demonstrated that *p*-values computed by pooling test statistics across genes tend to have a heavier tail than the true

Table 2. Real data analysis results: number of genes to estimate *p*-values (number of genes), estimated EE gene proportion ($\hat{\pi}_0$) and number of detected genes using *p*-value = 0.001, *q*-value = 0.01 and *q*-value = 0.05 from three microarray datasets having many (Data 1), moderate numbers of (Data 2) and few (Data 3) DE genes. (* out of 45101 total genes)

Data	Number of genes*	Procedure	$\hat{\pi}_0$	*p*-value 0.001	*q*-value 0.01	0.05
Data 1	29401	*tab.t*	0.44	1528	4983	11 105
		pool.t	0.47	409	3398	11 069
		pool.ts	0.47	441	3994	12 563
		poolb.t	0.48	1551	5707	12 655
		poolb.ts	0.48	3073	7308	13 663
Data 2	36904	*tab.t*	0.69	360	1573	4831
		pool.t	0.75	244	1612	5403
		pool.ts	0.74	281	1665	5350
		poolb.t	0.77	489	2176	6153
		poolb.ts	0.76	571	2253	6099
Data 3	38029	*tab.t*	0.78	110	865	3863
		pool.t	0.80	113	916	4098
		pool.ts	0.81	105	847	3927
		poolb.t	0.83	203	1368	4798
		poolb.ts	0.84	205	1283	4564

null distribution computed by permuting EE genes, and thus result in conservative inference. This is a consequence of the fact that the null distribution represents a mixture from EE and DE genes. Following Xie *et al*. (2005) and Fan *et al*. (2005), we proposed pooling using a subset of genes and demonstrated that such *p*-values can provide correct type I error, unbiased FDR estimates and good power. We recommend using the standard *t*-test to define the subset for pooling, but LOWESS curve fitting to the pooled *p*-values could be used to determine a threshold for subset selection. Our simulation study shows that $\alpha = 0.10$ critical value of the *t*-distribution serves well as a threshold in the situations studied here. The effects of the subset selection before permutation method are less pronounced when there are fewer DE genes, when the mean effect size is small and when the sample size is large (10 or more per group). For small experiments we found that complete enumeration of the permutation distribution was desirable and for larger experiments that no fewer than 1000 permutations should be used to obtain stable *p*-values.

The information borrowing statistics, t_s and t_b can be substantially more powerful than the standard *t*-test in small experiments. These two statistics show very similar performance. Selection of the subset for pooling should be done before computing these test statistics on the permuted data.

We have restricted attention to two condition comparisons using t, t_s and t_b statistics. However the method of subset selection and pooling extends directly to the case of multiple group comparisons. In this case we recommend using the standard *F*-test to select a subset and an information borrowing statistics such as F_s (Cui *et al*., 2005) or B statistics (Lonnstedt and Speed, 2002; Smyth, 2004) to carry out analysis. In the case of experiments with multiple sources of variation (random or mixed effects ANOVA) the F_s statistic allows fitting and shrinkage of multiple variance components. The subsetting before permutation method with F_s statistic is implemented in latest release of R/mannova (version 1.2.1 : http://www.jax.org/staff/churchill/labsite/software).

ACKNOWLEDGEMENTS

The authors thank Lei Wu and Qian Li for helpful discussion and testing software. This work was supported by NIH grant CA88327(G.C.).

Conflict of Interest: none declared.

REFERENCES

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc*., **74**, 829–836.

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Fan,J. *et al*. (2005) Removing intensity effects and identifying significant genes for Affymetrix arrays in macrophage migration inhibitory factor-suppressed neuroblastoma cells. *Proc. Natl Acad. Sci. USA*, **102**, 17751–17756.

Fisher,R.A. (1935) *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Ihaka,R. and Gentleman,R. (1996) A Language for data analysis and graphics. *J. Grap. Comput. Stat*., **5**, 299–314.

Lonnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sinica*, **12**, 31–46.

Smyth,G.K. *et al*. (2003) Statistical issues in cDNA microarray data analysis. *Meth. Mol. Biol*., **224**, 111–136.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, 1–26.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. Royal Stat. Soc*., **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: Parmigiani,G., Garrett,E.S., Irizarry,R.A. and Zeger,S.L. (eds), *The Analysis of Gene Expression Data: An Overview of Methods and Software*. Springer, New York, pp. 272–290.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat*., **31**, 2013–2035.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wu,H., Kerr,K. and Churchill,G.A. (2003) MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. *In The Analysis of Gene Expression Data: An Overview of Methods and Software*. Springer, New York, pp. 313–431.

Xie,Y. *et al*. (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, **21**, 4280–4288.