Comments on the analysis of unbalanced microarray data

Kathleen F. Kerr
Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA

Received on February 3, 2009; revised on May 20, 2009; accepted on June 9, 2009
Advance Access publication June 15, 2009
Associate Editor: Trey Ideker

ABSTRACT

Motivation: Permutation testing is very popular for analyzing microarray data to identify differentially expressed (DE) genes; estimating false discovery rates (FDRs) is a very popular way to address the inherent multiple testing problem. However, combining these approaches may be problematic when sample sizes are unequal.

Results: With unbalanced data, permutation tests may not be suitable because they do not test the hypothesis of interest. In addition, permutation tests can be biased. Using biased P-values to estimate the FDR can produce unacceptable bias in those estimates. Results also show that the approach of pooling permutation null distributions across genes can produce invalid P-values, since even non-DE genes can have different permutation null distributions. We encourage researchers to use statistics that have been shown to reliably discriminate DE genes, but caution that associated P-values may be either invalid, or a less-effective metric for discriminating DE genes.

Contact: katiek@uw.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

DNA microarray studies have become an integral part of biological research over the last decade. A common goal of microarray studies is to identify genes that are differentially expressed (DE) between two or more treatments or conditions. This is usually done in the framework of statistical hypothesis testing. Since the number of genes in a microarray study is typically in the tens of thousands, a severe multiple testing problem results (Dudoit et al., 2003).

Several trends have dominated common practice in analyzing microarray data in the past 5 years or so (Allison et al., 2006). We mention three trends here. First, the false discovery rate (FDR) is favored for addressing multiple testing. Second, classical test statistics (in particular, the t-statistic) have been found to work poorly in the microarray setting. Statistics developed specifically for microarrays do a better job at identifying DE genes. Third, permutation testing is very popular. We comment further on each of these in Sections 1.1–1.3, and discuss why methods that work well for balanced data may be problematic for unbalanced data.

1.1 False discovery rate

The classical method for addressing the multiple testing problem is to ensure that the probability of getting one or more false positives or ‘type I’ errors is small. This is called controlling the ‘familywise error rate’ (FWER). However, microarray investigators usually do not mind getting a small proportion of false positives as the trade-off for getting a sizable list of candidate genes for further study. As an alternative to the FWER, one can instead consider the proportion of type I errors among the rejected null hypotheses. Benjamini and Hochberg (1995) defined the FDR as FDR = E(V/R). R is the total number of rejected null hypotheses and V is the total number of type I errors among that set.

Allison et al. (2002) recognized that different methods for estimating the FDR are all based on treating the P-values from multiple hypothesis tests as a mixture distribution on the unit interval (0,1). Allison et al. (2002) call this family of techniques ‘Mixture Model Methods’ (MMM) for P-values. If the null hypothesis is true (a gene is not DE), then P(P-value ≤ α) = α, so the P-value is uniformly distributed on the unit interval. If the null hypothesis is false (a gene is DE), MMMs assume that, loosely speaking, the P-value distribution should have more mass near 0. MMMs have become very popular with microarrays and other high-throughput technologies; specific methods include Qvalue (Storey and Tibshirani, 2003), the Beta-Uniform Mixture Model (BUM) (Pounds and Morris, 2003), SPLSH (Pounds and Cheng, 2004), the Location Based Estimator (Dalmasso et al., 2005), the methods of Cheng et al. (2004), Do et al. (2005), Xie et al. (2005), Pounds and Cheng (2006), Heller and Qin (2007) and Ruppert et al. (2007).

Any P-value threshold α corresponds to a list of genes to be declared DE. MMMs are used to estimate the FDR associated with such a list. For a large number of P-values, E(V/ R) = α(1 - Fp(α)), where Fp is the cumulative distribution function (CDF) of all P-values.

It is simple and intuitive to use the observed number of P-values less than α to estimate Fp(α). Many, though not all, MMMs use this straightforward estimate of the denominator of (1). To estimate η0 for the numerator of (1), MMMs use the fact that P-values arising from the null hypothesis follow a U(0,1) distribution. Different MMMs invoke the U(0,1) distribution for null P-values in different ways to estimate η0. Note that the U(0,1) distribution is invoked twice in (1), first in estimating η0 and second in using E(V) = αη0. Finally, the FDR depends on the P-value threshold α but η0 does not. Therefore, it is convenient to compare MMMs by comparing their estimates of η0.

All MMMs make the following assumptions:

(I) When the null hypothesis is true, a P-value is distributed U(0,1).
The independent assumption (III) has been discussed at some length in the literature, both in papers proposing MMMs and elsewhere (e.g. Ge et al., 2003; Heller and Qin, 2007; Owen, 2005; Pawitan et al., 2006; Qiu and Yakovlev, 2006; Qiu et al., 2004, 2005). (I) is the assumption that P-values are valid. (II) is the assumption that the hypothesis test is unbiased. As we will see, with unbalanced sample sizes assumption (II) can be erroneous.

1.2 Microarray test statistics

Biologists originally used simple fold-change rules to select the genes declared to be DE in a microarray study. Statisticians emphasized the need to heed the formal components of inference. An obvious alternative to a fold-change rule is to use the two-sample t-test to identify DE genes. However, statisticians and biologists soon came to realize that the two-sample t-statistic

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

did not effectively discriminate DE genes. The denominator of the t-statistic is unstable with the sample sizes that are typical of microarray studies. With thousands of genes on a microarray, many exhibit small variability by chance, and often the genes with the largest t-statistics are those with the smallest denominators.

Alternative test statistics have been proposed (Cui et al., 2005; Smyth, 2004; Tusher et al., 2001) and are being used. In addition to the t-statistic, we will examine two alternative statistics: the \( \Delta \mu \) statistic,

\[ \Delta \mu = \bar{X}_1 - \bar{X}_2 \]

which is just the numerator of the t-statistic, and a 'SAM' statistic (Tusher et al., 2001),

\[ s = \frac{\bar{X}_1 - \bar{X}_2}{\frac{s_1^2}{\sqrt{n_1}} + \frac{s_2^2}{\sqrt{n_2}}} \]

where a constant \( \delta \) is added to the denominator of the t-statistic. We set \( \delta \) to be the median of all t-statistic denominators (see e.g. Xie et al., 2005). Statistics like the s-statistics are now largely favored in microarray studies because they excel at identifying the genes of greatest scientific interest.

To get a P-value from any test statistic one must know, theoretically or empirically, its distribution under the null hypothesis. Permutation appears to be the most popular technique for estimating null distributions with microarray data (Gao, 2006) and statistics like the s-statistic. Some of the implications of this are described next.

1.3 Permutation testing

The 'thought experiment' of a permutation test is that, under the null hypothesis, the sample labels are arbitrary and an alternate dataset with scrambled sample labels was just as likely as the observed dataset. Following the logic of this thought experiment, the null hypothesis of the test is the 'equality of distributions' hypothesis \( H_0: F_1 = F_2 \) (the \( F \)'s are the distributions of expression levels in the respective populations). This is a stronger null than the hypothesis that the two populations have the same means, which is tested, for example, by the two-sample t-test. See Dudouit et al. (2003) and Huang et al. (2006) for additional discussion of why \( H_0: F_1 = F_2 \) is the null hypothesis of a permutation test.

In our experience, when scientists are interested in DE genes, most often they are thinking of genes whose mean level of expression is different between groups. Translating this into statistical terms, the desired null hypothesis to test is the 'equality of means' hypothesis \( H_0: \mu_1 = \mu_2 \). This suggests that a permutation test is not suitable. However, even though a permutation test tests the 'equality of distributions' null, the differences in distributions that a particular test will have power to detect depend on the test statistic used and the sample sizes in the groups. In certain special cases, the test may only be sensitive to differences in the mean [see Huang et al. (2006), Pollard and van der Laan (2003) and Supplementary Appendix], making permutation testing a reasonable choice. However, in general a permutation test should be considered a test of the 'equality of distributions' null.

A related fact is that permutation tests can be biased, particularly when sample sizes are unbalanced. This means that, for example, a permutation test with size \( \alpha = 0.05 \) could potentially have power <0.05 for some alternatives. This 'negative power' is an obvious problem with such a test unless the test is only biased for alternatives that are not of interest. However, even if this special circumstance holds, the P-values from the biased test may not be suitable for multiple testing adjustment. For example, a biased test does not satisfy assumption (II) of MMMs. As we will see, the bias might only affect a minority of genes, but this can distort the distribution of P-values and throw-off methods that rely on the distribution of P-values across genes. Therefore, although permutation tests and MMMs are both popular and appealing, they are not always compatible for unbalanced data.

Supplementary Table 1 and Supplementary Figure 1 describe simple simulations that illustrate the kinds of P-value distributions that can arise from a permutation test. The permutation test used in those examples is biased whenever the sample size is larger in the sample with larger variance (see Supplementary Appendix).

In this article, \( H_0: \mu_1 = \mu_2 \) is considered to be the null hypothesis of interest. Certainly, there are situations where a different null hypothesis is appropriate (Ho et al., 2008).

2 METHODS

Microarray data were downloaded from ftp.sanger.ac.uk/pub/genevar/. The tissue is Epstein Barr Virus-transformed lymphoblastoid cell lines and samples come from 60 individuals with European ancestry (CEU) and 45 ethnic Chinese living in Beijing (CHB). The dataset used here is a version that was normalized all together and considered appropriate for joint analysis. There are data on 47 293 transcripts. We chose this dataset because it has fairly large sample sizes. Therefore, sample means and SDs should be reasonably accurate. We conduct simulations with large, intermediate and small sample sizes, in each case maintaining the 4:3 ratio of the original data.

For each gene and each group (CEU and CHB), we calculated the sample SD of the gene in that group. These values were then ‘coarsened’ to the
The QValue software is available at http://genomics.princeton.edu/storeylab/
We did this so that means could unambiguously be considered equal or
MMM assumptions. The EV simulations, therefore, provide a good basis for
/Δµ/
For each simulated dataset, a permutation test was performed using
2003) and CDE (Langaas et al. in Bioconductor (Gentleman
3 RESULTS
For each simulated dataset, a permutation test was performed using the Δµ-test statistic, \( X_{\text{CEU}} - X_{\text{CHB}} \). Figure 1 and Supplementary Figure 3 summarize the results of EV, UV1 and UV2 simulations. Consider first the EV simulations, where the MMM assumptions hold. We see that the MMMs are conservatively biased, as they are designed to be, tending to underestimate \( \pi_1 \). In general, as sample sizes increase, the bias and variability of the estimates of \( \pi_1 \) decrease, as one would expect, although some bias persists. See also Supplementary Figures 4 and 5 for the \( t \)-statistic.
Consider next the UV simulations, where genes are not forced to have the same variance in the simulated populations. Clearly, the MMM estimates are tending toward \( \pi_1^* \), not \( \pi_1 \). For the UV1 simulations, the permutation test is biased for the 13.4% of genes that have larger variance in the group with larger sample size (see Supplementary Appendix), which tends to decrease the estimates of \( \pi_1 \). In the UV2 simulations, the permutation test is biased for \( \sim t \) % of genes and has some power to detect differential expression for 13.4% of genes that are DE in the variance (regardless of whether they are DE in the mean). The MMMs therefore produce anti-conservative estimates of \( \pi_1^* \) for \( \pi_1^* \) of 1% and 5%. Of course, the MMM is supposed to estimate \( \pi_1 \), not \( \pi_1^* \), but the test is either biased or severely underpowered to detect genes DE in the variance. To be clear, the problem is not that the MMMs do not work. The problem is that the permutation test \( P \)-values in the UV simulations do not satisfy their assumptions. Histograms of \( P \)-values would not help a data analyst diagnose that there is a problem with the MMM assumptions (Supplementary Fig. 6).
When 20% or 45% of genes are DE in the mean (bottom two rows of Fig. 1), the difference between the type EV, UV1 and UV2 simulations is minimal. On the other hand, for smaller sample sizes, the bias in the estimates of \( \pi_1 \) is quite large. A statistician will recognize that the bias problem in the permutation test that was used here could be resolved by using a different test statistic. If we use the two-sample \( t \)-statistic instead of the Δµ-statistic, the null of the permutation test would have been the ‘equality of means’ hypothesis. This is probably the hypothesis biologists intend to test, and the permutation test would not be biased. However, as mentioned in Section 1.2, the \( t \)-statistic does a poor job in selecting DE genes. Figure 2 shows receiver operating characteristic (ROC) curves for three test statistics. The \( t \)-statistic does uniformly better than the \( \pi_1 \)-statistic at discriminating DE genes. The Δµ-statistic performs worse than the other two statistics at small false positive rates. (It actually performs best at higher false positive rates, but this is not the region of interest.) Figure 2 reinforces previous findings that the \( x \)-statistics (and statistics like it) do a superior job at discriminating DE genes (see, e.g. Cui et al., 2005; Qin et al., 2004; Smyth, 2004). A striking feature of Figure 2 is that the \( x \)-statistic does a better job at selecting DE genes than the \( \pi_1 \)-statistic \( P \)-values. (The ROC curves for the \( P \)-values are not shown in Figure 2, but all three are essentially the same as for the \( t \)-statistic.) This raises another important point. The permutation tests summarized in Figure 1 and Supplementary Figures 3-5 were ‘per gene’ permutation tests. That is, the data only for a particular gene were used to estimate the permutation null distribution of the test statistic. However, many microarray experiments do not have enough samples to get a permutation test \( P \)-value this way. For example, if the sample size is 4 in each of two groups, then the effective number of permutation is \( \binom{4}{2} = 70 \). Only a crude permutation density can be estimated with 70 datapoints, and the minimum \( P \)-value that can be obtained is \( 1/70 = 0.0014 \). One way to get around this problem is to combine the permutation nulls across all genes to get a single permutation null distribution. If the sample sizes are 4 and 4 and there are 20,000 genes, the ‘pooled’ permutation null is estimated by 70 * 20000 datapoints.
Fig. 1. QValue estimates of π₁ for the Δµ-statistic. A reference line is drawn at π₁, the proportion of genes DE in the mean. However, π₁ = π₁ only for the EV simulations. For the UV simulations, π₁ < π₁. CDE, an alternative MMM, gives similar results (Supplementary Fig. 3). See also Supplementary Figures 4 and 5 for the s-statistic.

Fig. 2. ROC curves for the Δµ-statistic (dashed line), s-statistic (solid line) and t-statistic (dotted line) in a type EV simulation with 10 000 genes. The ROC curves for the corresponding permutation test P-values are the same as for the t-statistic and are not shown.
Previous studies of the approach of pooling permutation nulls across genes have focused on the problem that DE genes will have different permutation null distributions than non-DE genes (Gao, 2006; Xie et al., 2005; Yang and Churchill, 2007). An apparently less recognized problem is that permutation null distributions can differ among non-DE genes (Fig. 3). When this is the case, P-values from a single reference null are prima facie invalid as permutation test P-values. However, obtaining P-values from a single reference distribution means that the rank ordering of the genes according to the P-values will be the same as the rank ordering of the genes according to the test statistic. The pooled permutation null distribution effectively induces a monotone transformation of the test statistic. Therefore, although using a single empirical null distribution may not produce valid P-values, these invalid P-values may do a superior job at selecting DE genes.

Figure 4 shows that 'per gene' permutation P-values and 'pooled null' permutation P-values differ substantially on the real CEU/CHB data. The most significant genes will be different depending on which kind of P-value is used. The difference between the two kinds of P-values is non-trivial.

Since the s-statistic provided the best discrimination of DE genes, it would be useful to have a way to estimate quantities such as the FDR directly from the statistic without needing P-values as intermediaries. Efron (2007) offers an approach that operates directly on test statistics rather than P-values. We investigated whether this approach offers a viable alternative to P-value MMs. We applied the associated software, ‘locfdr,’ to our simulations. By trying out different transformations of the s-statistics, we were able to get reasonable results (Fig. 5).

Specifically, we rescaled the statistics and transformed them using s-statistic P-values from a ‘per gene’ permutation test and a ‘pooled null’ permutation test. The inset magnifies the plot for small P-values. These results are for the real CEU/CHB data (not simulated data) with 10 000 randomly chosen genes to facilitate presentation. See Supplementary Figure 7 for the corresponding plots for other test statistics.
to small differences in transformations. For example, using a t-distribution that differed by just a few degrees of freedom dramatically affected results.

4 SUMMARY AND DISCUSSION

We are aware that, in practice, empirical P-value distributions sometimes do not conform to the kinds of distributions one expects (e.g. Fodor et al., 2007; Page et al., 2006). There are many possible explanations for such deviation, including large-scale dependencies among genes and important covariates that are not accounted for (Efron, 2007; Leek and Storey, 2007). Biased tests are another possible cause.

A common goal of microarray studies is to select genes that are DE in the mean. Data analysts therefore want to use a test statistic that discriminates DE and non-DE genes as effectively as possible, leading them to use a specialized test statistic. The null distribution of such a statistic must be estimated empirically, and data analysts often choose a permutation test to do this. Multiple testing is often addressed by applying an MMM to estimate the FDR for a list of genes called DE. This approach is problematic for unbalanced data. First, the permutation test is testing the ‘equality of distributions’ null, which is not usually the null hypothesis of interest. Second, the permutation test might be biased, in which case the P-values from the test are not appropriate to use with multiple-testing methods such as MMs. See Xu and Hsu (2007) and Calian et al. (2008) for a deeper exploration of issues that arise with permutation testing.

The s-statistic is one of many statistics designed specifically to discriminate DE genes in microarray data. The s-statistic and its cousins have been shown to perform better than t-statistics at this job, not just in the simulations presented here but in other simulations and empirical studies. However, we have seen that it can be problematic to rely on permutation testing to get valid, reliable P-values for the s-statistic. Our ‘per gene’ permutation test P-values were valid, but had two problems: (i) they were biased for some genes, and (ii) they were much less effective than the statistic at discriminating DE genes. P-values coming from a pooled permutation null distribution are invalid because different genes have different permutation null distributions. However, these ‘pseudo-P-values’ have the property of having a monotone relationship with the statistic, producing good results if used to select genes to call DE.

What, then, is the recommendation to the scientist who wants to identify DE genes? The overwhelming evidence favors using a variant of the t-statistic with a stabilized denominator, such as the s-statistic, to rank the genes. One option is to forego formal statistical inference and simply investigate the top genes according to such a discriminating test statistic. Independent confirmation is always more compelling than a small P-value. This answer is of course not very satisfactory when one wishes to publish results prior to independent confirmation, or to apply for the funding to do so. Another option is to use a method that does not rely on P-values such as ‘locfdr.’ However, we remain somewhat wary of the sensitivity of this method to choices that are made in fitting the model.

ACKNOWLEDGEMENTS

The author thanks Li-Xuan Qin and David Cobeltz for work on earlier projects that preceded this article, and Dick Beyer and the anonymous reviewers for numerous helpful suggestions.

Funding: Grant # NIEHS U19ES011387 to the FHCRC/UW Toxicogenomics Research Consortium; a grant from the National Heart Lung and Blood Institute # HL072370; Grant # NIEHS P30ES07033 to the UW Center for Ecogenetics and Environmental Health; Grant # NIEHS P50ES015915 to the UW Discover Center.

Conflict of Interest: none declared.

REFERENCES

Unbalanced microarray data


