BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 192–201
doi:10.1093/bioinformatics/btm583

Gene expression

Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis

Dan Nettleton[1,*], Justin Recknor[2] and James M. Reecy[3]

[1]Department of Statistics, Lowa State University, Ames, Lowa 50011-1210, USA, [2]Eli Lilly and Company, Lilly Research Laboratories, P.O. Box 708, Greenfield, Indiana 46140 and [3]Department of Animal Science, Lowa State University, Ames, Lowa 50011-3150, USA

Received on June 27, 2007; revised on October 11, 2007; accepted on November 19, 2007

Advance Access publication November 27, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The field of microarray data analysis is shifting emphasis from methods for identifying differentially expressed genes to methods for identifying differentially expressed gene categories. The latter approaches utilize a priori information about genes to group genes into categories and enhance the interpretation of experiments aimed at identifying expression differences across treatments. While almost all of the existing approaches for identifying differentially expressed gene categories are practically useful, they suffer from a variety of drawbacks. Perhaps most notably, many popular tools are based exclusively on gene-specific statistics that cannot detect many types of multivariate expression change.

Results: We have developed a nonparametric multivariate method for identifying gene categories whose multivariate expression distribution differs across two or more conditions. We illustrate our approach and compare its performance to several existing procedures via the analysis of a real data set and a unique data-based simulation study designed to capture the challenges and complexities of practical data analysis. We show that our method has good power for differentiating between differentially expressed and non-differentially expressed gene categories, and we utilize a resampling based strategy for controling the false discovery rate when testing multiple categories.

Availability: R code (www.r-project.org) for implementing our approach is available from the first author by request.

Contact: dnett@iastate.edu

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

1 INTRODUCTION

Microarray technology has enabled researchers to simultaneously measure the transcriptional expression levels of tens of thousands of genes in each of multiple biological samples. This information provided by microarray experiments is often used to identify genes that differ in expression across two or more treatments. Identification of such genes can provide clues about gene function and provide insight into the molecular genetic mechanisms underlying biological processes.

Deriving biological understanding from a list of genes that have been declared to be differentially expressed (DDE) is a challenging enterprise. Researchers typically use pre-existing information about the functions of genes to interpret the impact of varying treatments on gene expression. Pre-existing information about gene function can be derived from a variety of sources (Boeckmann *et al*., 2003; The Gene Ontology Consortium, 2000; Sonnhammer *et al*., 1997; Kanehisa and Goto, 2000). Regardless of the information source, the genes represented on a microarray slide can usually grouped into several categories, including a large category for genes with unknown function. The proportional representation of each category among the DDE genes can be compared to the proportional representation of each category on the microarray. Gene categories that are in some sense overrepresented or underrepresented among the DDE genes are then judged to play an important role for understanding how treatments affect the transcriptional program of the organism under study.

Many authors have proposed statistical methods for identifying gene categories that are over or underrepresented among DDE genes, and several software packages are available for scientists wishing to implement such analysis. References to much of the work in the area can be found in review articles by Khatri and Drăgichi (2005) and Allison *et al*. (2006). The most popular approaches use variations of Fisher's exact test to identify categories whose representation among the DDE genes differs significantly from the expected representation under an often unstated null hypothesis. Basically, such procedures test whether the number of DDE genes from a certain category is significantly more or less than would have been expected if the DDE genes had been randomly and independently drawn without replacement from the collection of all genes represented on the microarray.

While such tests are intuitively appealing to many scientists as a natural way to focus attention on categories with many (or few) DDE genes, the methods have been criticized on statistical grounds by Allison *et al*. (2006), Barry *et al*. (2005), Subramanian *et al*. (2005), and Goeman and Bühlmann (2007) among others. We review some of these criticisms and

*To whom correspondence should be addressed.

add a few of our own in Section 3. Partly in response to the drawbacks associated with these methods, Subramanian *et al.* (2005) and Barry *et al.* (2005) derived new methods for identifying gene categories of interest. Although the methods differ considerably in approach, they share the common goal of attempting to identify gene categories that consist of genes that are in some sense significantly more differentially expressed than all other genes.

Rather than pitting groups of genes against each other to compete for attention based on gene-specific assessments of differential expression, we propose a nonparametric multivariate method for identifying gene categories whose joint expression distribution differs across treatments. We view our approach as a more powerful and more natural method for identifying gene categories of special interest. As simple motivation for our multivariate approach, consider Figure 1 which shows—for each of two treatments—the multivariate expression distribution for a category consisting of two genes. Although the multivariate expression distribution of the category clearly depends on the treatment, the distribution of each individual gene is identical for both treatments. Thus, regardless of sample size, the treatment effect illustrated in Figure 1 would be invisible to the existing methods that are based on measuring differences across treatments separately for each gene. In contrast, the method we propose would easily identify the depicted two-gene category as a category of interest with sufficiently large samples for each treatment group.

Although the example in Figure 1 is completely hypothetical, we find that our method exhibits greater levels of discovery for real data relative to comparable methods based on gene-specific tests. We illustrate this behavior using one example in Section 4. Our example involves a microarray experiment aimed at understanding expression differences between wild type mice and mutant mice lacking a functional copy of the Myostatin gene, an inhibitor of skeletal muscle growth. The mutant mice develop muscles that weigh 2–3 times those of the wild type mice. We show how to use our methodology to identify groups of genes involved in creating this dramatic phenotype. The full details of our method are described in Section 2. Section 3 contrasts our method with existing approaches. A simulation study is described in Section 5. Concluding remarks are provided in Section 6.

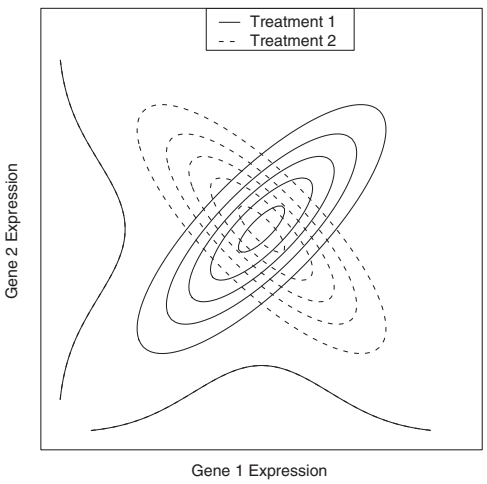

Fig. 1. Distinct joint and non-distinguishable marginal expression distributions for two treatments and a hypothetical gene category consisting of two genes.

2 MODEL

2.1 Testing a single gene category for differential expression across two or more treatments

Let T denote the number of treatments, n_i denote the number of independent replications of treatment i ($i=1,\ldots,T$), and G denote the number of genes in a category of interest. For $i=1,\ldots,T$ and $j=1,\ldots,n_i$; let $\boldsymbol{Y}_{ij}=(Y_{ij1},\ldots,Y_{ijG})'$ denote a vector of expression measurements, where Y_{ijk} represents the expression measurement associated with treatment i, replication j, and gene k. We assume that all $\boldsymbol{Y}_{ij}$ vectors are independent and that $\boldsymbol{Y}_{ij}$ has a continuous multivariate distribution F_i. We wish to test

$$H_0 : F_1 = \cdots = F_T \tag{1}$$

against all alternatives. When this null hypothesis is false, the multivariate distribution of the genes in the category of interest is not the same for all treatments. For a completely randomized experimental design, violation of H_0 implies that at least one treatment caused a change in the category's multivariate expression distribution. Thus, categories for which H_0 is false are of potential scientific interest.

To test H_0 against all alternatives, we propose to use the multiresponse permutation procedure (MRPP) described by Mielke and Berry (2001). The MRPP test statistic is given by

$$\bar{D} = \sum_{i=1}^{T} \frac{n_i}{N} D_i, \tag{2}$$

where $N = \sum_{i=1}^{T} n_i$ and D_i is the average of all the Euclidean distances between pairs of data vectors from the ith treatment group, i.e.

$$D_i = \frac{\sum_{j=1}^{n_i-1} \sum_{j'=j+1}^{n_i} ||\, Y_{ij} - Y_{ij'} ||}{n_i(n_i-1)/2}. \tag{3}$$

The MRPP test uses a standard permutation approach to assess the significance of the observed value of $\bar{D}$. The permutation p-value is given by

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\bar{D} \geq \bar{D}_m), \tag{4}$$

where $\bar{D}_m$ denotes the value of the test statistic computed for the mth of $M = N!/\prod_{i=1}^{T} n_i!$ possible assignments of treatment labels to the observed data vectors, and $\mathbb{1}(\cdot)$ denotes the indicator function that takes the value 1 if its argument is satisfied and 0 otherwise. As with any permutation test, if the number of data permutations M is too large for timely computation, a randomly selected subset of permutations can be used to obtain an approximate permutation P-value.

Extensive justification of the MRPP test is provided by Mielke and Berry (2001). Here we briefly note that, conditional on the observed data vectors, $\bar{D}$ will be relatively small when the data vectors associated with different treatments are well separated in G-dimensional space. Thus, the testing procedure can detect location shifts in the multivariate expression distribution caused by treatment effects. However, the MRPP tests also has power to detect departures from H_0 other than location shifts. For example, the MRPP test yielded an average P-value of 0.035 over 100 simulations of samples of size 30 from the distributions depicted in Figure 1. Thus, the MRPP test can detect evidence against H_0 even when both treatment distributions have identical means, substantial overlap, and differences that are invisible to marginal approaches. Of course, it is possible to devise other test statistics that have greater power for detecting the departure to H_0 depicted in Figure 1 (e.g. the difference between correlation coefficients), but the strength of the MRPP test lies in its ability to detect a wide variety of departures from H_0 using a single, relatively simple procedure.

2.2 Accounting for Variance Differences Among Genes

It is well known that different genes exhibit different levels of variation. If this heterogeneity of variance is not accounted for, genes with larger variability can dominate the results of our proposed MRPP test. This situation is depicted in Figure 2(A).

The plot shows bivariate expression observations for 10 experimental units from each of two treatment groups. Data were simulated so that the within-treatment SD of gene 1 was 10-fold less than the within-treatment SD of gene 2. The bivariate means were chosen so that the treatments have identical means with respect to gene 2 but different means with respect to gene 1. Thus, the null hypothesis in (1) is false. However, the average distance between data vectors within a treatment group is quite high because these distances are dominated by variation in gene 2. As a consequence, the permutation P-value for our test of H_0 is approximately 0.28, and we would fail to detect what appears to be an obvious departure from H_0.

Mielke and Berry (2001) proposed an approach referred to as Euclidean commensuration for adjusting for heterogeneity of variance like that depicted in Figure 2(A). In our context, the data from the kth gene are scaled by

$$\left\{\sum_{i=1}^{T}\sum_{i'=i}^{T}\sum_{j=1}^{n_i}\sum_{j'=j}^{n_{i'}}(Y_{ijk}-Y_{i'j'k})^2\right\}^{-1/2} \tag{5}$$

prior to analysis, which is equivalent to standardizing that data for each gene to a common variance. After scaling, the sum of the squares of all possible pairwise within-gene differences is identical for all genes. Figure 2(B) shows the example data after commensuration. The MRPP P-value is approximately 0.0001 when working with the scaled data. Thus, the departure from H_0 is easily detected following commensuration. Note that commensuration will not remove heterogeneity of variance across treatments within a gene. Thus, potentially interesting changes in variance across treatments may still be detected following commensuration.

In the example presented here, commensuration aided in the identification of multivariate differential expression. However, in other circumstances, power for detecting multivariate differential expression may be greater for non-commensurated data. It is well known that variation of expression differs substantially from gene to gene. When using non-commensurated data, the MRPP test statistic effectively gives greater weight to genes with higher variation. If the higher variation genes happen to contain the essential information about treatment differences, emphasis on high variation genes will be well placed, and power for detecting differences will be greater when using non-commensurated as opposed to commensurated data. Goeman *et al.* (2004) argue that a strength of their proposed global test procedure is that genes with large variance have much more influence on their test statistic than low-variability genes. We will demonstrate via data analysis and simulations in Sections 4 and 5, respectively, that the MRPP approach with non-commensurated data is very similar to the global test method. Henceforth, we will use MRPP_C to denote the MRPP approach with commensurated data.

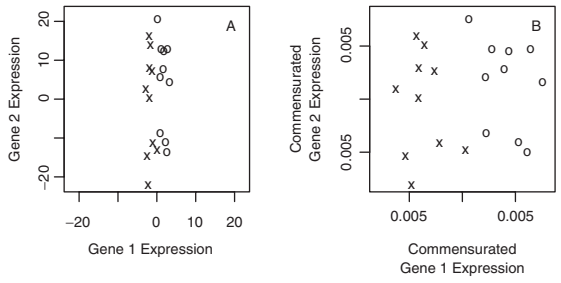

Fig. 2. Bivariate expression data for two treatments before and after Euclidean commensuration. Treatment 1 (2) observations are marked with x (o). Within-treatment SD of gene 1 is 10-fold less than within-treatment SD of gene 2 prior to commensuration.

2.3 Simultaneous testing of multiple categories

The MRPP approach can also be used to test each of multiple gene categories for evidence of differential expression across treatments. The categories cannot be assumed to be independent because genes from different categories are not necessarily independent of one another and because single genes may be associated with more than one category. GO annotations, for example, group genes into categories with varying levels of specificity. The union of several specific categories will form a subset of the genes in a more general category. Simultaneously testing several categories of varying levels of specificity can be useful for pinpointing the nature of treatment effects on the expression program of an organism. Thus, we need to employ a multiple testing procedure that remains effective when tests are dependent.

Following Barry *et al.* (2005), we will use a version of the resampling-based false discovery rate (FDR) controling procedure developed and studied by Yekutieli and Benjamini (1999) and Reiner *et al.* (2003). To describe this procedure in our context, some additional notation is required. Let $\bar{D}^c$ denote

the MRPP statistic $\bar{D}$ in (2) computed for the cth of C categories. Let $\bar{D}^c_m$ denote the value of $\bar{D}^c$ computed using the mth permutation of the treatment labels relative to the experimental units. We adopt the convention that permutation 1 represents the original assignment of treatment labels to the experimental units; thus, $\bar{D}^c = \bar{D}^c_1$ for all $c = 1, \ldots, C$. Now for any $c = 1, \ldots, C$ and $m = 1, \ldots, M$; we define

$$p^c_m = \frac{1}{M}\sum_{m'=1}^{M} \mathbb{1}(\bar{D}^c_m \geq \bar{D}^c_{m'}).$$

Note that p^c_1 is the permutation P-value for testing for treatment effects on the multivariate expression distribution of genes in category c as presented in (4) in Section 2.1. For a given threshold for significance p, we declare all categories with $p^c_1 \leq p$ to be differentially expressed. The estimate of FDR associated with the selected threshold for significance is given by

$$\widehat{\mathrm{FDR}}(p) = \min_{p':p' \geq p}\left(\frac{1}{M-1}\sum_{m=2}^{M}\frac{R_m(p')}{R_m(p') + S(p')}\right), \qquad (6)$$

where

$$R_m(p') = \sum_{c=1}^{C} \mathbb{1}(p^c_m \leq p')$$

and

$$S(p') = R_1(p') - \frac{1}{M-1}\sum_{m=2}^{M} R_m(p').$$

Reiner *et al*. (2003) refer to $\widehat{\mathrm{FDR}}(p)$ as the 'resampling-based point estimator' of FDR. Yekutieli and Benjamini (1999) developed this estimator for controlling FDR when conducting multiple dependent tests. Simulations in Yekutieli and Benjamini (1999) and Reiner *et al*. (2003) establish good power and FDR control characteristics even when test statistics are highly correlated. We illustrate the use of this procedure in Section 4 and investigate its performance in the gene category testing context via simulation in Section 5.

3 COMPARISON WITH EXISTING APPROACHES

Our proposed procedure is an alternative to approaches that compare the categorical composition of a list of DDE genes to the categorical composition of genes on a microarray using testing procedures like Fisher's exact test (e.g. Berriz *et al*., 2003; Drăghici *et al*., 2003; Doniger *et al*., 2003; Al-Shahrour *et al*., 2004; Beibarth and Speed, 2004; Cheng *et al*., 2004). In most such methods, a category is considered to be enriched among DDE genes if the number of DDE genes in the category is larger than would have been expected had the DDE genes been a simple random sample from all genes on the microarray. Of course, prior to conducting such a test for enrichment, it is well known the that DDE genes differ from a simple random sample of genes on a microarray. In particular, genes are not independent of one another, so simple random sampling provides a poor probability model for the selection of DDE genes. What appears to be an unusually large (or small) number of genes from a given category under an assumption of simple random sampling might be easily explained by positive correlation among genes in a category rather than 'enrichment'. Goeman and Bühlmann (2007) refer to such methods for computing P-values as *gene sampling* methods and use simulation to show that such approaches can be quite liberal when genes are dependent.

Beyond the question of statistical validity, several authors (Barry *et al*., 2005; Subramanian *et al*., 2005; Allison *et al*., 2006) have pointed out that information is lost when continuous gene-specific measures of differential expression (e.g. P-values) are dichotomized to produce DDE and non DDE genes. The rank order of the evidence for differential expression within DDE and non-DDE gene lists is lost. Furthermore, the results of such enrichment analyses can be sensitive to the threshold for significance used to produce the DDE genes. In some cases, all genes in a given category may exhibit small changes that, when considered together, provide strong evidence of a treatment effect. However, such categories will go undetected if many of the individual changes fail to reach the chosen threshold for significant differential expression. Newton *et al*. (2007) have developed methods for addressing this criticism; however, their approach uses a *gene sampling* probability model and is thus suspect when genes are dependent.

Gene Set Enrichment Analysis (GSEA) (Subramanian *et al*., 2005) and Significance Analysis of Function and Expression (SAFE) (Barry *et al*., 2005) are two approaches that address many of the deficiencies of previous methods. Both GSEA and SAFE begin by computing a measure of differential expression for each gene. GSEA and SAFE each use a different measure of differential expression, but we will simply refer to either measure generically as a gene statistic. Next, the gene statistics for a given category are compared to the gene statistics from all genes outside the given category. The comparison is summarized with a category statistic. Again, the computational details differ between the methods, but the main point is that a category statistic is produced (one for each category of interest) that measures the extent to which the genes in a given category are more differentially expressed than genes outside the given category. Both methods then identify significant categories by comparing the observed category statistics from the data at hand to the distribution of category statistics that is obtained by permuting the treatment labels relative to the observed data vectors. Goeman and Bühlmann (2007) call this a *subject sampling* approach for determining significance as opposed to the problematic *gene sampling* approach discussed previously.

Neither Subramanian *et al*. (2005) nor Barry *et al*. (2005) explicitly state the null hypothesis tested for each gene category when using GSEA or SAFE, respectively. Based on the construction of the GSEA and SAFE category statistics, it might be natural to assume that these methods are testing a null hypothesis of 'no enrichment' for each category. However, their permutation testing procedures are not justified under these null hypotheses because they permute treatment labels relative to data vectors (*subject sampling*) rather than permuting gene labels relative to gene statistics (*gene sampling*).

Furthermore, it would be quite difficult to develop testing procedures of the 'no enrichment null' because of correlation among genes (Efron and Tibshirani (2007): Goeman and Bühlmann (2007) for excellent additional discussion of this issue). The validity of the GSEA and SAFE permutation testing procedures is most easily justified under the complete null hypothesis that the joint expression distribution of all genes is identical for all treatments. Such a null is of limited interest, however, and it would more desirable and natural to consider separate category-specific nulls of the form

$$H_0^{(c)} : F_1^{(c)} = \cdots = F_T^{(c)}$$

where $F_i^{(c)}$ denotes the joint expression distribution of genes in category c under treatment i. These are exactly the null hypotheses that we consider in our proposed approach; i.e. we consider simultaneously testing $H_0^{(1)}, \ldots, H_0^{(C)}$ for $c = 1, \ldots, C$.

To obtain strong control of error rates in this multiple testing problem (i.e. control regardless of which subset of the tested null hypotheses are true), the subset pivotality condition introduced by Westfall and Young (1993) and assumed by Yekutieli and Benjamini (1999) is required. However, this condition does not hold for either the GSEA or SAFE approaches. To clarify this point, let P_c denote, for any method, the P-value for testing $H_0^{(c)}$. Subset pivotality dictates that the joint distribution of $\{P_c: c \in S\}$ for any subset of categories S is the same under $\bigcap_{c \in S} H_0^{(c)}$ as under $\bigcap_{c=1}^{C} H_0^{(c)}$. Because each category statistic in the GSEA and SAFE approaches is obtained by comparing the gene statistics in a given category to the gene statistics for all genes outside the category, the subset pivotality condition is clearly violated. We do not encounter this difficulty with our approach because our statistic for a given category is a function of only the genes in the given category. In the terminology of Goeman and Bühlmann (2007), our test is *self contained* as opposed to *competitive*.

Comparison of gene statistics within a category to gene statistics outside a category (*competitive testing*) can also be problematic for less technical reasons. The presence of categories with many substantially differentially expressed genes will make it more difficult for GSEA and SAFE approaches to detect other categories whose expression patterns have been affected by treatment. This is potentially a desirable characteristic for researchers who wish to focus on only a few prominent categories. However, researchers wishing to understand the full impact of treatments on the expression pattern of an organism might gain greater insight from our approach. Because categories do not compete with each other for attention, more extensive discovery of differentially expressed categories is possible.

As previously noted, a strength of our approach is the ability to detect treatment effects on the multivariate expression distribution of genes in a category. In contrast, the SAFE approach, the GSEA approach, and the improvements to the GSEA approach proposed recently by Efron and Tibshirani (2007) can detect only treatment effects on gene-specific marginal distributions. Tomfohr *et al.* (2005), Goeman *et al.* (2004), and Liu *et al.* (2007) have proposed methods that can detect multivariate changes in joint expression distributions that may not be easily detected when focusing only on marginal distributions. The methods are similar in spirit to our approach, but the category statistics used to detect changes in the expression distribution are quite different.

The Pathway Level Analysis of Gene Expression (PLAGE) method of Tomfohr *et al.* (2005) involves the computation of the 'activity level' of the genes in a given category as a linear combination of the expression measures of genes in the category. The coefficients of the linear combination are the components of the leading eigenvector in the singular decomposition of the expression sub-matrix corresponding to the genes in the category of interest. This is equivalent to computing the first principal component scores for each experimental unit, separately within each category. Two-sample t-statistics using these activity levels as data are proposed as category statistics in the two-sample problem. We expect this to be a useful method for testing $H_0^{(c)}$ in most cases. However, the first principal component will not always be a good summary of the data for detecting differences across treatments. For the data in Figure 1, for example, the mean of the first principal component will be nearly identical for both treatments. Thus, a two-sample t-test will fail to detect a treatment effect. We compare the performance of PLAGE with our MRPP approach for the analysis of a real data set in Section 4 and in a simulation study in Section 5.

Liu *et al.* (2007) recently proposed Domain Enhance Analysis (DEA) strategies that include DEA principal component analysis (DEA-PCA) and DEA partial least squares analysis (DEA-PLS). DEA-PCA is equivalent to the PLAGE approach of Tomfohr *et al.* (2005). The DEA-PLS method is a new and promising approach that may share many of the benefits of our multivariate method. There are, however, some drawbacks to the current implementation that we note below. Rather than summarizing the expression vectors with the first principal component, the first partial least squares component is used. In this case, the first partial least squares component is the linear combination of the expression values within a category that has maximum covariance with numerically coded treatment labels. For the two-treatment case, Liu *et al.* use a 0/1 coding but acknowledge that different codings (e.g. $1/-1$) result in different test statistics with different properties. Alternative approaches that explicitly account for the categorical nature of the treatment labels were mentioned by the authors, but to our knowledge, none of these strategies have been implemented. Once the first PLS component has been computed for the 0/1 coding, a standard two-sample t-test with Benjamini and Hochberg's (1995) adjustment for multiple testing is used to identify differentially expressed categories. Unfortunately, this testing procedure is not valid because, as noted by Liu *et al.* the null distribution of the t-test statistic will have heavier tails that the usual t distribution due to construction of the first PLS component as the linear combination that has maximum covariance with the numerically coded treatment labels. This leads to type I error rates and false discovery rates that are higher than nominal. To address this problem, it may be necessary to develop a permutation strategy as described in Section 2.3. However, such a strategy would be considerably more computationally intensive than the corresponding strategy for the PLAGE method because, unlike

the first principal component that remains unchanged across permutations of the treatment labels, the first partial least squares component must be recomputed for each permutation.

The Global Test (GT) method of Goeman *et al*. (2004) addresses the slightly different problem of identifying gene categories whose expression can be used to predict a clinical outcome (e.g. cancer versus no cancer). The clinical outcome is modeled as the response variable in a generalized linear mixed model. The expression of genes in a category of interest are used as explanatory variables in the generalized linear mixed model, and the coefficients on these explanatory variables are assumed to be normally distributed with mean 0 and variance τ^2. A test of $H_0 : \tau^2 = 0$ is conducted separately for each category. If H_0 is rejected for a given category, there is evidence that the clinical outcome depends on expression of genes in the category. The method could be applied to the two-treatment problem that we consider in this article by viewing treatment levels as the 'clinical outcome.' The method could also be extended to handle the case of more than two treatments. We investigate the performance of this approach on a real dataset in the next section and in simulation in Section 5.

4 APPLICATION TO A MYOSTATIN KNOCKOUT EXPERIMENT

Myostatin is a protein that inhibits the rate of muscular cell growth and differentiation. Cattle with mutations in the gene responsible for myostatin production, such as the Belgian Blue and Piedmontese, have increased quantities of muscle mass. Understanding what other proteins are affected by the suppression of myostatin can provide insight into the molecular mechanisms underlying muscle. Mutant mice that had their myostatin gene knocked out were studied in a recent experiment by Steelman *et al*. (2006) to understand differences in gene expression caused by myostatin production. The experiment compared the expression levels of the mutant mice to that of wild-type mice at three time points. The time points were selected to test the impact of myostatin at both the primary and secondary stages of muscular tissue formation along with a time of fast muscular growth. At each time point, five experimental units per genotype where each measured using the Affymetrix GeneChip Mouse Expression Set 430.

One of the interests of the study was to find gene categories defined by GO terms which exhibited expression differences between the mutant and wild-type mice at the time of fast muscular growth. For this investigation, the GO molecular function terms associated with each Affymetrix probe set were obtained from the Bioconductor moe430a package. Of the 22 690 probe sets included in the Affymetrix GeneChip Mouse Expression Set 430 A Chip, 18 565 were assigned to at least one of the GO molecular function categories tested. A total of 353 molecular function categories included 40 or more probe sets. Following Barry *et al*. (2005), we tested only these 353 categories (though our method is applicable to categories of any size).

With five experimental units for each of the two genotypes at the time point of interest, there are $\binom{10}{5} = 252$ possible permutations. For a two-sided permutation test, the smallest possible *P*-value is 1/126. A total of 77 of the 353 categories tested had the minimum *P*-value of 1/126 when applying MRPP_C approach. The null hypothesis was rejected for these categories at a false discovery rate of $\approx 2.5\%$, which was estimated using the approach proposed by Yekutieli and Benjamini (1999) as described in Section 2.3. When using non-commensurated data, 22 categories were detected by the MRPP method as significantly differentially expressed with a *P*-value of 1/126 and an estimated FDR of $\approx 8.1\%$.

Table 1. Number of categories declared to be differentially expressed for each method (diagonal entries) and number of categories declared to be differentially expressed by both of two methods (off-diagonal entries) for the myostatin data set

Method	MRPP_C	MRPP	PLAGE	GT	SAFE	GSEA
MRPP_C	77	7	35	6	0	4
MRPP	–	22	7	19	0	0
PLAGE	–	–	55	7	0	1
GT	–	–	–	20	0	0
SAFE	–	–	–	–	0	0
GSEA	–	–	–	–	–	12

The same 353 GO molecular function categories were also tested using the GSEA, SAFE, PLAGE and GT methods described in Section 3. To allow for comparison of the methods on equal terms, permutation *P*-values were computed for these other methods as described above for the MRPP approach. Table 1 displays the number of categories having permutation *P*-values equal to 1/126 for each method along with the number of categories meeting that criterion for both methods. The MRPP approach with commensuration declared the most categories to be differentially expressed and typically offered far more unique discoveries than any single competing method. In fact, MRPP_C identified 36 categories as differentially expressed that were not detected by any other method.

Note that GSEA and SAFE discovered far fewer categories than PLAGE, GT and either MRPP procedure. This is not surprising because MRPP, PLAGE and GT are multivariate procedures that use *self-contained* testing with *subject sampling* to assess significance as recommended by Goeman and Bülmann (2007). In contrast, the GSEA and SAFE procedures are neither multivariate nor *self contained* and—as noted by Goeman and Bülmann (2007)—are almost invariably less powerful than *self-contained* testing procedures. Although SAFE identified no categories as significant at the significance level 1/126 compared to 12 found by GSEA, we have examined other data sets not presented here where SAFE declared more categories to be differentially expressed than GSEA. Thus, performance of GSEA and SAFE on this data set is not indicative of the general performance of these two procedures.

Also of note is the close agreement between the MRPP approach without commensuration and the GT method. These approaches identified 19 categories in common among the 23 categories identified by one or both methods. The rank

correlation between the *P*-values of the two methods across the 353 categories was 0.970. Thus, although these testing procedures appear quite different on the surface, they captured nearly the same information about differential expression in this data set.

The MRPP approach with commensuration identified 77 categories compared to 22 categories detected by the MRPP approach with non-commensurated data. Thus, in this case, commensuration appears to have aided the discovery of differential expression substantially. Of course, this is an analysis of a single data set for which the truth about differential expression is unknown. Therefore, we cannot rule out the possibility of false discovery. However, we will demonstrate in the next section that our FDR estimation method tends to work well for both versions of the MRPP approach which suggests that the results presented here are trustworthy.

To gain some insight into the strong performance of the MRPP approach with commensuration, we will focus on the category associated with the GO molecular function term 'nucleotidyltransferase activity.' This category consisting of 174 probe sets was detected as differentially expressed only when using $MRRP_C$. Conducting a two-sample *t*-test for each gene and applying the method of Nettleton *et al*. (2006) suggests that approximately 15 of the 174 nucleotidyltransferase activity genes were differentially expressed between mutant and wild-type mice. However, none of the gene-specific changes were particularly dramatic (minimum *P*-value > 0.001). The MRPP approach, however, can identify differences in the multivariate distribution that are invisible to the *t*-test approach. Examination of a scatter plot of the first two principal components of the commensurated data reveals that the mutant and wild-type mice are well separated in this two-dimensional space (see Supplementary Figure 1). If Fisher's linear discriminant is computed using this two dimensional data, we obtain a linear combination of the standardized data that dramatically separates the mutant and wild-type mice (linear combination values of 10.4, 11.8, 1.9, 6.9, and 6.7 for mutant mice compared to −13.7, −8.7, −6.2, −5.4, and −3.6 for wild-type mice). This linear combination tends to place positive (negative) weight on genes for which mutant mice exhibited higher (lower) levels of expression than wild-type mice. The rank correlation between the *t*-statistics and the weights in the linear combination was 0.938.

The first principal component utilized by PLAGE does not show separation between mutant and wild-type mice, so it is not surprising that PLAGE failed to detect a difference for this category (PLAGE *P*-value ≈ 0.111). The GT method and the MRPP approach without commensuration yielded *P*-values of 0.325 and 0.349, respectively. Examination of the principal components for non-commensurated (or, equivalently, non-standardized) data shows that the mutant and wild-type mice are not well separated by any linear combination of the first two principal components. Thus, information about differential expression is masked when data are not commensurated in this case. However, our simulation study discussed in the next section shows that commensuration will not always aid in the identification of multivariate differential expression.

5 SIMULATION STUDY

While the performance of the MRPP approach on the myostatin data set is encouraging, the results in Section 4 do not prove the superiority of our method because the true differential expression status of the gene categories is unknown in the myostatin experiment. In this section, we examine the performance of the MRPP approach in a unique data-based simulation study that allows us to assess the power and error control properties of our procedure and to further compare its performance with the other multivariate methods PLAGE and GT.

We based our simulation on the B- and T-cell Acute Lymphocytic Leukemia (ALL) data set described in part by Chiaretti *et al*. (2004) and analyzed in gene category testing studies by Liu *et al*. (2007) and Jiang and Gentleman (2007). The data set is publicly available in the Bioconductor ALL package at www.bioconductor.org. The data consist of 12 625-dimensional expression profiles from the Affymetrix HGU95aV2 GeneChip for each of 128 ALL patients. Of the 128 patients, 95 suffer from B-cell ALL while 33 have T-cell ALL. Using version 1.16.0 of the hgu95av2 Bioconductor package, we were able to map 10 467 of the Affymetrix probe sets to at least one GO term from the biological processes ontology, and 4153 terms from the biological processes ontology were each associated with at least one probe set.

Liu *et al*. (2007) analyzed the ALL data to identify the most significantly differentially expressed categories (defined by biological processes ontology terms) for their DEA-PLS method and the Fisher's exact test approach. The top 10 categories for each method are described in Tables 4 and 5 of Liu *et al*. (2007). We chose 15 categories from the union of these two top ten lists to serve as differentially expressed categories in our simulation study. These 15 categories involve 1274 of the 12 625 probe sets in the ALL data. Additional information about the selected categories is provided in Supplementary Table 1.

The following procedure was used to generate each of 100 simulated data sets.

(1) Randomly select (without replacement) $2n$ of the 95 B-cell samples and randomly divide the selected samples into two 'treatment' groups of n samples each.

(2) Create two 12 635 by n data matrices (one for each group) from the 12 625-dimensional expression vectors associated with the selected B-cell samples.

(3) For each of n T-cell samples randomly selected without replacement from the 33 T-cell samples, extract the 1274-dimensional expression sub-vector corresponding to the probe sets associated with the differentially expressed categories in Supplementary Table 1.

(4) For each column of one of the matrices created in step 2, replace the 1274-dimensional expression sub-vector corresponding to the probe sets associated with the differentially expressed categories with one of the T-cell expression sub-vectors randomly selected in step 3.

Provided that the categories in Supplementary Table 1 are indeed differentially expressed between B- and T-cell samples

(a safe assumption given the results in Liu *et al*.), the simulation procedure generates differential expression between the two treatments for each category in Supplementary Table 1, and no differential expression between treatments for each category that does not involve the 1274 probe sets associated with the Supplementary Table 1 categories. There are 1848 such categories among the 4153 categories that we considered. The other 2290 (=4153-15-1848) categories each involve one or more genes from the 1274 probe sets associated with Supplementary Table 1 categories. Note that these categories may or may not be differentially expressed in our simulation procedure because not all of the 1274 probe sets associated with Supplementary Table 1 categories are necessarily differentially expressed. Thus, we focused on the ability of the MRPP approach to correctly distinguish the differentially expressed categories in Supplementary Table 1 from subsets of the 1,848 categories that are guaranteed to be non-differentially expressed based on our simulation procedure. In particular, for each simulated data set, we applied PLAGE, GT and both MRPP approaches to the 15 categories in Supplementary Table 1 and 85 categories randomly selected from the 1848 categories that are guaranteed to be null by our simulation design. We chose to randomly select 85 null categories for each simulated data set—rather than analyzing all 1848 null categories—to ease the computational burden in our simulation and to match our belief that 15 differentially expressed categories out of 100 is a more realistic differential expression fraction than 15 of 1863.

Note that we could have searched for differentially expressed categories other than those identified by Liu *et al*. (2007) by analyzing the full ALL data set using our MRPP approach. However, that could have biased our simulation study to favor our MRPP approach because the types of differentially expressed categories identified would have been those whose differential expression characteristics made them relatively easy to detect with the MRPP approach. To avoid this potential bias, we chose to analyze categories selected by methods distinct from those considered in our simulation study. Furthermore, we focused only on the very most significant categories identified by Liu *et al*. (2007) to avoid the possibility of including categories that are not truly differentially expressed between B- and T-cell ALL samples.

Of course, the purpose of our data-based simulation strategy is to mimic, as closely as possible, the types of multivariate differential expression that can occur in an actual microarray data set. Traditional simulation strategies would require the complete specification of multivariate distributions for each treatment. Such specifications may not realistically represent actual multivariate expression distributions, correlations among genes and gene categories, the nature of multivariate differential expression across treatments, etc. Thus, we have used random sampling within a real data set to create a simulation study that should have greater practical relevance than traditional studies that are farther from real data structures. Note that, as is often the case with real data sets, there is likely to be substantial heterogeneity within the B- and T-cell classes (Yeoh *et al*., 2002). Thus, methods that work well in our simulation need to be relatively robust to within-treatment heterogeneity that is common in large microarray data sets.

Table 2. Mean and SEM across 100 simulated data sets for the quantities $V =$ number of false positives, $R =$ number of rejected null hypotheses, $\frac{V}{R \vee 1} =$ observed false positive fraction, and $\widehat{\text{FDR}} =$ the estimated false discovery rate computed using the method described in Section 2.3 for $n = 15$ observations per treatment and significance threshold P-value ≤ 0.01

Method		V	R	$\frac{V}{R \vee 1}$	$\widehat{\text{FDR}}$
MRPP_C	mean	1.32	8.77	0.078	0.087
	SEM	0.53	0.64	0.016	0.005
MRPP	mean	0.77	9.94	0.058	0.075
	SEM	0.18	0.23	0.011	0.001
GT	mean	0.78	9.61	0.062	0.078
	SEM	0.18	0.23	0.011	0.001
PLAGE	mean	0.73	1.63	0.147	0.559
	SEM	0.28	0.35	0.032	0.035

We considered two sample sizes ($n = 5$ and 15) and three thresholds for significance (P-value = 0.01, 1/126 and 5/126). For the sample size $n = 15$ case, 999 randomly selected data permutations along with the observed sample were utilized to compute permutation P-values for the methods. All 252 data permutation were analyzed for the $n = 5$ case. 100 simulated data sets were generated for each scenario. The results for $n = 15$ and significance threshold 0.01 are reported in Table 2. Results for other scenarios are qualitatively identical and are reported in Supplementary Tables 2 and 3.

The MRPP procedure without commensuration was the top performing method in the simulation study. MRPP had the greatest number of discoveries on average and the lowest average false positive fraction among all the competing methods. The GT method performed nearly as well as MRPP. Although not evident from the means and standard errors in the Table 2, the performance of the MRPP procedure was significantly better than GT. For example, the two procedures correctly identified precisely the same differentially expressed categories for 59 of the 100 simulation replications. For the other 41 replications where there was some disagreement between the methods, the MRPP approach identified more true positives than GT 35 times, GT identified more true positives than MRPP 5 times, and the methods identified the same number of true positives (though not the same categories) once. A paired-data sign test (McNemar's test) yields significance at well below the 0.0001 level. Though statistically significant, the differences between the MRPP and GT procedures were small, and as would be expected given the results in Section 4, the rank correlation between their P-values across all simulated data sets was quite high (0.944).

MRPP_C and PLAGE trailed MRPP and GT in this simulation. Unlike MRPP and GT, both MRPP_C and PLAGE operate on standardized data. In contrast to the myostatin data analyzed in Section 4, standardization seems to hamper the ability to detect differential expression in the ALL data set. If principal components are computed for the 1274

probe sets associated with the Supplementary Table 1 categories using non-standardized data, there is a dramatic separation between B- and T-cell samples in a plot of second versus first principal components (see Supplementary Figure 2). On the other hand, if data are standardized (a typical practice when computing principal components), the B- and T-cell samples are well mixed in the analogous principal component plot (see Supplementary Figure 3). This illustrates that much of the information about differential expression in the ALL data set is contained in higher variation genes. Hence, the methods that use standardized data ($MRPP_C$ and PLAGE) were outperformed by the methods that use non-standardized data (MRPP and GT). Note that among the methods using standardized data, $MRPP_C$ was substantially better than PLAGE. Though the mean number of false discoveries was slightly higher for $MRPP_C$, the number of true positives ($R-V$) was much greater for the $MRPP_C$ method. Consequently, the average observed false positive fraction—which serves as an empirical estimate of FDR—was much lower for the $MRPP_C$ approach.

Note that for all methods, the estimated FDR levels were, on average, greater than the empirical estimates, indicating that FDR estimation procedure discussed in Section 2.3 performed conservatively in this study. The results reported in Table 2 are specific to the 0.01 threshold for significance and sample size of $n=15$ per treatment, but as mentioned previously and illustrated in Supplementary Tables 2 and 3, the same basic conclusions can be drawn when considering different thresholds for significance (P-value $\leq 1/126, 5/126$) and smaller sample sizes ($n=5$).

6 EXTENSIONS

We have restricted our presentation to completely randomized designs where an overall test of distributional equality as in (1) is of interest. Many microarray experiments have factorial treatment structures and/or multiple sources of variation that should be accounted for in analysis. Several books discuss the use of permutation testing methods in such circumstances. Mielke and Berry (2001) cover MRPP methods for factorial experiments and experiments that include multiple sources of variability (e.g. split-plot designs). However, their discussion of complex designs is confined mostly to the univariate case. Pesarin (2001) presents multivariate permutation testing methods with an emphasis on combining univariate tests. Edgington (1995) presents extensive material on the analysis of factorial experiments, multivariate analysis and multivariate analysis of factorial experiments in the context of randomization tests. Much of this material is directly relevant to our approach and can be used as a guide for conducting MRPP tests for experimental designs that are more complex than we have considered here. For example, randomized complete block designs are easily handled by ignoring blocks in the computation of the MRPP statistic and computing the permutation P-value using only those data permutations that involve exchanging treatment labels among experimental units within blocks. Testing for the effects of a single factor in a factorial experiment can be accomplished in a similar manner. See chapters 6 through 8 of Edgington (1995) for more details and additional discussion of other complex designs.

7 CONCLUSION

We have proposed a non-parametric multivariate method for identifying differentially expressed gene categories. Rather than testing for differences in enrichment between categories, we have focused on testing for treatment differences within categories. The strength of our method is the ability to detect general changes in each category's multivariate expression distribution that would be invisible to the many popular methods that rely on functions of gene-specific statistics. We cannot claim that the MRPP test upon which our method is based is the best of all possible multivariate tests for identifying differentially expressed gene categories, but it has clearly performed very well in our studies. In particular, MRPP performed as well or better than leading multivariate procedures when multivariate treatment distributions differed for high-variation genes, and $MRPP_C$ appeared to perform the best of all the competing procedures when data standardization aided in identifying multivariate differential expression. Although' no one multivariate testing procedure will be superior to all others for all scenarios, we believe the MRPP testing strategy will provide excellent performance across a wide range of practical situations. Our simulation work shows that our method can distinguish differentially expressed from non-differentially expressed categories while providing estimates of the false discovery rate that are not overly optimistic. Reducing the apparent conservativeness of our FDR estimates is a topic worthy of future research.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for comments that improved this manuscript. This material is based upon work supported by the National Science Foundation under Grants No. 0500461 and 0714978.

Conflict of Interest: none declared.

REFERENCE

Al-Shahrour,F. *et al.* (2004) Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature*, **7**, 55–65.

Barry,W. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Beibarth,T. and Speed,T. (2004) Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini,Y. and Hochberg,Y. (1995) Controlling false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Berriz,G. *et al.* (2003) Characterizing gene sets with funcassociate. *Bioinformatics*, **19**, 2502–2504.

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Cheng,J. *et al.* (2004) NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.

Chiaretti,S. *et al*. (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Doniger,S.W. *et al*. (2003) MAPPFinder: using gene ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*., **4**, 7.

Drăghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Edgington,E.S. (1995) *Randomization Tests*. 3rd edn. Marcel Dekker, New York.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat*., **1**, 107–129.

The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Goeman,J.J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27–30.

Khatri,P. and Drăghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Liu,J. *et al*. (2007) Domain enhanced analysis of microarray data using GO annotations. *Bioinformatics*, **23**, 1225–1234.

Mielke,P., Jr. and Berry,K. (2001) *Permutation methods: A Distance Function Approach*. Springer-Verlag, New York.

Nettleton,D. *et al*. (2006) Estimating the number of true null hypotheses from a histogram of p-values. *J. Agri., Bio., Environ. Stat*., **11**, 337–356.

Newton,M.A. *et al*. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat*., **1**, 85–106.

Pesarin,F. (2001) *Multivariate Permutation Tests with Applications in Biostatistics*. Wiley, Chichester.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Sonnhammer,E.L.L. *et al*. (1997) Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. *Proteins*, **28**, 405–420.

Steelman,C. *et al*. (2006) Transcriptional profiling of myostatin-knockout mice implicates Wnt signaling in postnatal skeletal muscle growth and hypertrophy. *FASEB J*., **20**, 580–582.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci*., **102**, 15545–15550.

Tomfohr,J. *et al*. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan Inference*, **82**, 171–196.

Yeoh,E.-J. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.