Bioinformatics Advance Access published February 11, 2009

A Flexible Two Stage Procedure for Identifying Gene Sets that are Differentially Expressed

Ruth Heller [1,*], Elisabetta Manduchi [2], Gregory R. Grant [2] and Warren J. Ewens [3*]

[1]Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340
[2]Computational Biology and Informatics Laboratory, Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104-6021
[3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Motivation: Microarray data analysis has expanded from testing individual genes for differential expression to testing gene sets for differential expression. The tests at the gene set level may focus on multivariate expression changes or on the differential expression of at least one gene in the gene set. These tests may be powerful at detecting subtle changes in expression, but findings at the gene set level need to be examined further to understand whether they are informative and if so how.

Results: We propose to first test for differential expression at the gene set level but then proceed to test for differential expression of individual genes within discovered gene sets. We introduce the overall FDR (OFDR) as an appropriate error rate to control when testing multiple gene sets and genes. We illustrate the advantage of this procedure over procedures that only test gene sets or individual genes.

Availability: R code (www.r-project.org) for implementing our approach is included as supplementary material.

Contact: ruheller@wharton.upenn.edu

1 INTRODUCTION

In recent years microarray data analysis has expanded from considering individual genes as units of interest to include gene sets as units of interest (see Nam and Kim (2008) for a review). A gene set is a collection of genes typically defined by prior knowledge about biochemical pathways, biological processes, or co-expression in previous studies. For example, the gene ontology (GO) has been used to define gene sets based on three criteria: (1) biological process, (2) molecular function and (3) cellular component (GO Consortium (2000), `http://www.geneontology.org`).

Many statistical methods have been proposed for testing the differential expression of gene sets. The methods have been classified in Tian et al. (2005) and in Goeman and Buhlmann (2007) in terms of the type of the gene set null hypothesis being tested. The first hypothesis type, termed Q1 or the *competitive null*, compares the gene set to a standard defined by the complement of that gene set. The second hypothesis type, termed Q2 or the *self-contained null*, compares the gene set to a fixed standard that does not depend on the measurements of genes outside the gene set. Excellent discussions of the advantages and disadvantages of each hypothesis type can be found in Goeman and Buhlmann (2007) and Nettleton et al. (2008). In short, self-contained hypotheses will typically be more powerful and have a clear biological meaning.

We are going to restrict our attention to self-contained hypotheses. In the category of self-contained hypotheses, there are many ways to define what is meant by differential expression of a gene set between a control group and a treatment group. Two natural definitions are: (1) that the joint distribution of expression levels for genes in the gene set is different between the control group and the treatment group (Nettleton et al. (2008)); (2) that at least one of the genes in the gene set is differentially expressed - this is a test of the global null hypothesis (Goeman et al. (2004), Goeman and Buhlmann (2007)). These characterizations have the advantage of being very general, so a test may be quite powerful. However, the disadvantages are respectively that (1) the joint distributions can differ in subtle ways that can be hard to interpret; and (2) having only one gene of a set differentially expressed may not be very meaningful biologically.

Therefore, a reasonable approach is to start by using one of the two characterizations above for initial screening. We call the null hypothesis of this first test a *screening hypothesis*. The choice of the screening hypothesis may depend on what the gene sets represent (e.g. biological processes or cellular components). However, as pointed out above, the finding that the screening hypothesis is false may be too vague and possibly not scientifically meaningful. Therefore, if the screening hypothesis is rejected for a gene set, this gene set needs to be examined further to understand whether the gene set finding is informative, and if so how. A natural follow-up question is to identify which genes in the gene set are differentially expressed, if any. This follow-up question has been considered in the widely used GSEA (Subramanian et al. (2005)). Subramanian et al. (2005) suggested the *leading-edge subset* analysis, that extracts the core members of the gene set based on an intuitive yet ad-hoc threshold. Specifically, the leading-edge subset (as defined in Subramanian et al. (2005)) consists of the genes in the gene set

*to whom correspondence should be addressed

that appear in the ranked list at, or before, the point where the running sum (used for calculating the enrichment score) reaches its maximum deviation from zero. Subramanian et al. (2005) answered the follow-up question of which genes are differentially expressed in an *exploratory* fashion, i.e. without controlling an error rate. We suggest instead a *confirmatory* strategy, i.e. one that controls an error rate. Confirmatory analysis is necessary to be able to answer whether the gene set findings are likely to be reproducible. Having done the confirmatory analysis, exploratory analysis can further be performed to generate more questions or hypotheses on how the gene set is differentially expressed, and these hypotheses may be validated with new data.

The approach we describe above involves first a test of a screening hypothesis for each gene set, and second a collection of additional tests for the gene sets with rejected screening hypotheses, so we have multiple hypotheses per gene set. We call the two stages in our approach the screening stage and the confirmation stage respectively. For the screening stage, a natural error measure is the expected proportion of gene sets for which the null hypothesis was incorrectly rejected, out of all gene sets for which the null hypothesis was rejected, that is the false discovery rate (FDR, see Reiner et al. (2003) for a review). We suggest the use of the *overall FDR* (OFDR), introduced in Benjamini and Heller (2008), as an appropriate and tractable error measure for the two stage procedure. Let a *discovered gene set* be a gene set for which the screening hypothesis has been rejected, and let a *falsely discovered gene set* be a discovered gene set for which at least one null hypothesis (including possibly the screening hypothesis) was incorrectly rejected. The OFDR is the expected proportion of falsely discovered gene sets out of all discovered gene sets.

Another possible error measure when multiple hypotheses are tested per gene set may be to control the FDR on all hypotheses for all gene sets. To illustrate the difference between the two error measures, consider the following example. Let the screening hypothesis be that none of the genes in the gene set are differentially expressed, so a falsely discovered gene set is a gene set with at least one falsely discovered gene. Suppose each of 1000 gene sets contains 100 genes to be examined for differential expression. Suppose 100 genes are discovered to be differentially expressed, and these genes belong to 20 gene sets. Suppose further that 25 of these genes have been falsely discovered, and all the falsely discovered genes belong to the same gene set. Then the proportion of false gene discoveries is $\frac{25}{100} = 25\%$, but the proportion of falsely discovered gene sets is $\frac{1}{20} = 5\%$. This toy example suggests that by controlling the OFDR, the FDR on the level of individual genes is not controlled. Similarly, controlling the FDR on the level of individual genes does not guarantee control of the OFDR. To see this, suppose that 5 genes are falsely discovered, and each of these genes belongs to a different gene set. Then the proportion of false gene discoveries is $\frac{5}{100} = 5\%$, but the proportion of falsely discovered gene sets is $\frac{5}{20} = 25\%$. For testing multiple hypotheses on gene sets, controlling the OFDR has two main advantages: (1) since the inferential units of interest are the gene sets, the error measure should control the false discoveries at the gene set level; and (2) the multiplicity problem may be less severe for an OFDR controlling procedure. To see this note that, in the example, an OFDR controlling procedure considers only the 1000 gene sets as the units for inference, but an FDR controlling procedure considers the individual genes as the units for inference and there are usually more genes than gene sets. The 1000 initial screening hypotheses may be more powerful tests than the tests on the individual gene hypotheses since for testing the screening hypothesis the evidence of 100 genes is pooled together. Moreover, the number of screening hypotheses is smaller than the number of individual gene hypotheses. Therefore, the OFDR controlling procedure may be more powerful than the FDR controlling procedure. See Sections 4 and 5 for a comparison of the performance of these two procedures on real and simulated microarray data respectively.

In Section 2 we describe a testing strategy for one gene set. In Section 3 we describe a general procedure that controls the OFDR when multiple gene sets are examined. In Section 4 we give an example from microarray data analysis that demonstrates the advantages of our suggested procedure over existing methods. In Section 5 we evaluate the OFDR controlling procedure by simulation and compare it to the leading-edge subset analysis in GSEA. In Section 6 we give some final remarks.

2 A TESTING STRATEGY FOR ONE GENE SET

Our proposed testing strategy for one gene set consists of two steps: (1) test a screening hypothesis, H_{screen}; and (2) if the screening hypothesis is rejected, test a collection of additional null hypotheses $H_1, \ldots, H_n$ on that set (e.g. test for differential expression on single genes one by one).

For step (1), a test that compares the joint expression of genes across groups has to be performed. Tests that are targeted towards detecting multivariate changes in joint expression distributions have been proposed by Goeman et al. (2004), Liu et al. (2007), and Nettleton et al. (2008), among others. We choose the nonparametric test introduced in Nettleton et al. (2008) as our screening hypothesis because it has good power for detecting subtle changes in expression between two groups within a gene set. The test statistic is computed as follows: let D_i be the average of all Euclidean distances between pairs of data vectors from group $i \in \{1, 2\}$; then the test statistic is $\overline{D} = \frac{m}{m+n} D_1 + \frac{n}{m+n} D_2$ for m cases in the first group and n cases in the second group. The test uses a standard permutation approach to assess the significance of the observed $\overline{D}$ (see Nettleton et al. (2008) for details). The smaller the observed $\overline{D}$ is compared to the values of this statistic for permuted samples, the stronger the evidence that the groups have different (multivariate) expression distributions. Let p_{screen} be the p-value for testing H_{screen}. For step (2), let the unadjusted p-values be $p_1, \ldots, p_n$. For example, these p-values may be the Wilcoxon rank sum test unadjusted p-values for comparing the expression levels of two groups for each gene.

If step (1) is performed at level α, and step (2) is performed with family-wise error rate (FWER, the probability of making at least one incorrect rejection of a null hypothesis) control at level α, then the FWER on $H_{\text{screen}}, H_1, \ldots, H_n$ is controlled at level α. To see that this procedure controls the FWER at level α, note that if the screening hypothesis is true then $\text{FWER} = Pr(P_{\text{screen}} \leq \alpha) \leq \alpha$, and if the screening hypothesis is false then the FWER is the probability of rejecting at least one of $H_1, \ldots, H_n$ falsely and this probability is at most α. This procedure is a "Gatekeeper" method (Bauer et al. (1998)), in which tests at the individual gene level are performed only if the test on the gene set level is significant. The advantage of the "Gatekeeper" method is that the test at the gene set level need not be adjusted for multiplicity.

Common stepwise procedures to control the FWER on $H_1, \ldots, H_n$ are (1) the Bonferroni-Holm procedure (Holm (1979)) which can be applied under any general dependency between the p-values; and (2) the more powerful Hochberg procedure (Hochberg (1988)) which can be applied if the p-values are independent. If $H_i, i = 1, \ldots, n$ correspond to null hypotheses on individual genes, the stepwise procedure may have little power if n is large. In fact, the stepwise procedure may fail to reject any of the individual gene hypotheses since the procedure relies on the evidence from individual gene p-values only (whereas the screening hypothesis combines the individual gene p-values).

Ge et al. (2003) show that for comparing a control group and a treatment group for differential expression on each gene, the step-down procedure in Westfall and Young (1993) based on minP adjusted p-values is a powerful procedure that makes minimal assumptions. The strength of this procedure is that it exploits the dependencies in the test statistics in order to improve power. Therefore, if the group sizes are not too small these step-down procedures are recommended. Implementations of these procedures are available in the R bioconductor package `multtest` (Pollard et al. (2008)).

When multiple gene sets are simultaneously examined, the above testing strategy has to be adjusted and this is done in Section 3.

3 THE CONTROL OF THE OVERALL FDR

Suppose that for each gene set $s \in \{1, \ldots, S\}$ we have a screening hypothesis $H_{\text{screen}}(s)$ and additional hypotheses $H_1(s), \ldots, H_{n(s)}(s)$. For controlling the FDR when only one hypothesis is tested per gene set, Benjamini and Hochberg (1995) suggested the first and still very popular FDR controlling procedure, called the BH procedure. When more than one hypothesis is tested per gene set, we suggest the following procedure and show that it controls the OFDR:

PROCEDURE 3.1. *The hierarchical testing procedure:*

1. *Screening Stage:*
   - *Choose the screening hypothesis for each set* $s \in \{1, \ldots, S\}$.
   - *Let* $\{p_{\text{screen}}(s) : s = 1, \ldots, S\}$ *be the unadjusted* p*-values for the screening hypotheses.*
   - *Apply the BH procedure at level* q *to these* p*-values. Let* R *be the number of rejected screening hypotheses.*
2. *Confirmation Stage: For each rejected set* s*,*
   - *compute the adjusted* p*-values for each gene in the gene set (e.g. using the Westfall and Young (1993) step-down minP procedure) .*
   - *Test these adjusted* p*-values at level* Rq/S*, rejecting the gene hypotheses with adjusted* p*-values* $\leq Rq/S$.

An estimated lower bound for the proportion of genes differentially expressed in set s is $u(s)/n(s)$, where $u(s)$ is the number of rejected gene hypotheses in set s, and since the OFDR is controlled at level q (see below) we expect at most q of these lower bounds to be false.

THEOREM 3.1. *Procedure 3.1 controls the OFDR at level* q*, assuming that the* p*-values of each gene set are independent from all other screening* p*-values.*

See Appendix 7.1 for a proof.

The independence assumption is unrealistic in gene set analysis. However, previous works have shown that the BH procedure controls the FDR for quite general dependencies among the p-values. Benjamini and Yekutieli (2001) proved that the BH procedure controls the FDR for positive regression dependency on each test statistic corresponding to a true null hypotheses (PRDS). See Benjamini and Yekutieli (2001) for examples where such dependency arises. Reiner (2007) argued via a combination of simulations and analytic results that applying the BH procedure on p-values from two-sided tests of correlated normal test statistics with any correlation structure controls the FDR. Storey et al. (2004) gave convergence conditions on the p-values for the BH procedure to control the FDR asymptotically. Reiner et al. (2003) show in simulations that for typical dependency between the test statistics in microarray data the BH procedure controls the FDR on individual genes. Similarly, for gene set analysis we show in simulations in Section 5 that for typical microarray dependency the FDR is controlled when applying the BH procedure to the screening hypotheses p-values. The simulations in Section 5 also show that Procedure 3.1 controls the OFDR under realistic microarray dependencies. In most practical situations where the BH procedure is appropriate, the OFDR controlling procedure will also be appropriate. If in doubt the more conservative level of $q/(\sum_{j=1}^{S} \frac{1}{j})$ can be applied to guarantee FDR control at level q at the gene set level (Benjamini and Yekutieli (2001)).

Finally, a cautionary remark about the use of other FDR controlling procedures instead of BH in the hierarchical testing procedure 3.1. The BH procedure is known to be conservative by the proportion of null hypotheses tested, π_0, so that even when the p-values are independent the FDR is controlled at the level $\pi_0 q$ (Benjamini and Hochberg (1995)). This conservativeness motivated the development of procedures that first estimate the number of null hypotheses and then use the estimate to enhance power (Benjamini et al. (2006), Storey et al. (2004), Ge et al. (2003)). However, applying these methods at the screening stage of Procedure 3.1 instead of the BH procedure may be nonconservative and the OFDR may not be controlled. To see this, note that if in every gene set at least one true null hypotheses is tested, then Procedure 3.1 may control the OFDR exactly at level q. Thus Procedure 3.1 may not be conservative in such a setting. However, if there are S_1 gene sets for which all null hypotheses considered are false, then Procedure 3.1 is conservative by the factor $\frac{S-S_1}{S}$. If S_1 was known, then performing in Procedure 3.1 the screening stage at level $\frac{S-S_1}{S}$ (instead of at level q) and the confirmation stage at level $\frac{Rq}{S-S_1}$ (instead of at level $\frac{Rq}{S}$) will be more powerful than Procedure 3.1 and still control the OFDR at level q.

4 APPLICATION TO A MICROARRAY STUDY

We use the data set of Chiaretti et al. (2004), available in the Bioconductor ALL package at `www.bioconductor.org`. The

Table 1. The distribution of the number of individual gene discoveries within the 418 discovered gene sets. ADD GENE NAMES.

#genes discovered	0	1	2	3	4	5	6	7	9	10	12	14
#gene sets	198	138	67	38	18	5	7	9	2	1	1	1

data set was collected to identify genes that distinguish subgroups of leukemia patients. It consists of 128 patients split into 95 with B-cell and 33 with T-cell type acute lymphoblastic leukemia (ALL). The data consist of 12,625 expression profiles from the HGU95aV2 Affymetrix chip for each patient. Using version 2.0.1 of the hgu95av2 Bioconductor package, we were able to map 9671 of the Affymetrix probe sets to at least one GO term for biological process, and 4364 terms from the GO biological process terms were each associated with at least one probe set. After restricting our gene sets to be of minimum size 2 and maximum size 500, we had 3367 gene sets to examine. The union of the 3367 gene sets included 8678 individual probes.

4.1 Method

We performed the following analysis for a random sample of 10, 15, or 20 replicates per group.

We applied Procedure 3.1 at level $q = 0.05$: the screening p-values were calculated by the method in Nettleton et al. (2008) (using 5000 permutations of the group labels to compute the gene set p-values); the FWER adjusted two-sided p-values were computed by the Westfall and Young (1993) step-down minP procedure using the Wilcoxon test statistic. We refer to this procedure as the *screening+minP* procedure.

For comparison, we applied the GSEA method. Gene sets with adjusted FDR below 0.05 where considered discovered gene sets, and the leading edge analysis was performed only on the discovered gene sets.

4.2 Results

We first examine the results of applying screening+ minP procedure at the 5% level on the data with 10 replicates per group. We discovered 418 gene sets and 87 unique genes within the discovered gene sets. The number of individual gene discoveries per gene set ranged from 0 to 15, see Table 1 for the distribution. The gene sets with 10, 12 and 14 individual gene discoveries were of size 220, 24 and 70 respectively.

Table 2 shows a sample of discovered gene sets and discovered genes within these gene sets. This sample corresponds to the discovered gene sets with the largest proportion of individual gene discoveries in the set. By applying the screening+minP procedure we expect at most 5% of the gene set discoveries to be false discoveries (i.e. the FDR is controlled on the screening hypothesis). In addition, by testing the individual genes we expect at most 5% of the rows in the complete results table of 418 (formatted as Table 2) to contain false positives (i.e. the OFDR is controlled). Note that if only one gene is discovered, it may be that there is only one differentially expressed gene in this set and a more satisfactory explanation for this gene being significant is that it participates in another GO process. On the other hand, discovering a large proportion of genes in a gene set gives strong evidence that this GO process behaves differently between the two groups being investigated. In this example, we sorted the rows by decreasing order of proportion of individual genes discovered within discovered gene sets with rejected screening hypotheses, to help focus attention to the gene set discoveries where most of the individual genes are differentially expressed. It may also be useful to look at the sorted rows by decreasing order of the number of individual genes discovered, or by a score that combines the number and proportion of individual gene discoveries in discovered gene sets. For example, a discovered gene set of size 20 with 15 individual gene discoveries may be more relevant to the researcher than a discovered gene set of size 2 with 2 individual gene discoveries.

Table 2. For the ALL data with 10 type B replicates and 10 type T replicates, the discovered gene sets, discovered probe set IDs (gene symbols), and percent of probe discoveries in the gene set, for the gene sets with the largest proportion of gene discoveries within the gene sets using the screening+minP procedure at the 5% level. Expect only 5% of the rows in the complete table of 418 rows to contain false positives.

Discovered Gene Sets	Discovered Probes Sets (Genes)	Percent Discovered
GO:0046827 positive regulation of protein export from nucleus	1253_at (GSK3B), 40645_at (GSK3B)	100
GO:0043368 positive T cell selection, GO:0045059 positive thymic T cell selection	1498_at (ZAP70), 38319_at (CD3D)	100
GO:0050862 positive regulation of T cell receptor signaling pathway	2059_s_at (LCK), 33039_at (TRAT1), 33238_at (LCK)	75
GO:0046825 regulation of protein export from nucleus	1253_at (GSK3B), 40645_at (GSK3B)	67
GO:0006426 glycyl-tRNA aminoacylation	36581_at (GARS), 36582_at	67
GO:0045022 early endosome to late endosome transport	41164_at (IGHM), 41165_at, 41166_at (IGHM)	60

Table 3 shows the number of gene set discoveries and the number of unique gene discoveries after applying the screening+minP procedure to data with 10, 15 or 20 replicates per groups. As the number of replicates increases, so does the number of discoveries. The number of gene set discoveries with a group size of 15 (1086) is much larger than that with a group size of 10 (418) but only slightly smaller than with a group size of 20 (1229). Similarly, the number of individual gene discoveries with a group size of 15 (300) is much larger than that with a group size of 10 (87) but only a little smaller than with a group size of 20 (383).

The GSEA method did not discover any gene sets with groups of size 10 or 15, and only 4 gene sets with groups of size 20. The leading-edge subset analysis identifies 30 individual genes.

Finally, note that in a situation where many individual genes are differentially expressed, a gene by gene analysis may find more individual genes than the proposed method. but the interpretability of the discoveries by the proposed method is greater since they can be associated with biological processes. In a gene by gene analysis it is not clear how to associate the list of discovered genes to gene

Table 3. For data with 10, 15, or 20 replicates per group, the number of gene set discoveries and the number of unique probe discoveries after applying the screening+minP procedure.

# Replicates per group	# Gene sets discovered	# Unique probes discovered
10	418	87
15	1086	300
20	1229	383

sets while controlling the FDR on individual genes (see Goeman and Buhlmann (2007) for a critique of available methods).

5 A SIMULATION STUDY

The goals of these simulations are (1) to verify that the Procedure 3.1 controls the OFDR for microarray dependency structures; and (2) to compare this procedure to that of a gene-by-gene analysis as well as to a variant of Procedure 3.1 that controls the FDR instead of the FWER when testing after the screening stage (see details in Section 5.1.2). Note that FDR control for each gene set does not necessarily imply FDR control for the entire study (Benjamini and Yekutieli (2005)), nor does it imply OFDR control. Note moreover that OFDR control of Procedure 3.1 implies FDR control on the screening hypotheses, thus showing that the BH procedure controls the FDR for microarray dependence in gene set analysis.

5.1 Simulation Study Design

5.1.1 Data Based Simulation Model In order to capture the complex dependence among genes and gene sets in microarray data, we performed the following data based-simulation, similar to that suggested in Nettleton et al. (2008).

The procedure below was used to generate 40 simulated data sets from the data described in Section 4:

1. Randomly select (without replacement) 30 of the 95 B-cell replicates and randomly divide the selected samples into two treatment groups of size 15 each.
2. Create two 12625 by 15 data matrices (one for each group) from the 12625 dimensional expression vectors associated with the selected B-cell replicates.
3. Extract the genes identified by the following 10 GO biological process terms: GO:0002263, GO:0019882, GO:0019883, GO:0019884, GO:0019885, GO:0019886, GO:0030097, GO:0045058, GO:0046649, GO:0050862. 345 unique genes were in the union of these 10 gene sets. We introduced differential expression into these genes by adding a constant $\Delta \in \{0.1, 0.2, 0.3, 0.4\}$ to the expression of these genes for each case in the first group.

Three remarks follow. First, note that the correlation between pairs of genes is unchanged within each group by the artificial introduction of differential expression. Second, since the gene sets are not mutually exclusive, 1649 gene sets contain differentially expressed genes even though we introduced differential expression only into the 345 genes that are in the union of the 10 gene sets listed in 3. Third, since the standard deviations of the expression levels vary across genes (they range from 0.1 to 2.3, with a median of 0.3 and an inter-quartile range of [0.2,0.5]), the power to detect the same difference of size Δ varies across genes.

5.1.2 Microarray Data Analysis Procedures We applied the two procedures in Section 4.1: screening+minP and GSEA.

We also did a typical gene-by-gene analysis on the 8678 individual genes: first we computed the unadjusted Wilcoxon p-values for individual genes; on these p-values we applied the BH procedure at level 0.05. We refer to this procedure as the *gene-by-gene analysis* procedure. Two remarks follow about the choice of this gene-by-gene analysis. First, note that the BH procedure is just one of many gene-by-gene analysis procedures that control the FDR. In this simulation setting, it is conservative by the factor $\pi_0 = \frac{8678-345}{8678} = 0.96$, which is close to 1. Therefore, the power of the BH procedure is expected to be similar to that of procedures that first estimate π_0 (discussed in Section 3). Second, we only consider an FDR controlling procedure since Reiner et al. (2003) showed that FDR controlling procedures obtain substantially more power than FWER controlling procedures.

In addition, we applied two FDR controlling procedure after the screening stage instead of the minP procedure for FWER control. Yekutieli (2008) suggested two level hierarchical FDR procedures in another context. Although there is no guarantee of controlling an error measure for gene set analysis using the screening+FDR procedures, they are intuitively appealing since FDR controlling procedures may be more powerful than FWER controlling procedures. We wanted to see how much power we lose by the fact that the FWER is controlled instead of the FDR within detected gene sets.

The first procedure is the BH procedure at level $\frac{R}{S}0.05$, where R is the number of rejected sets and S is the number of gene sets (S=3367 in our simulation). The second procedure is the minP-augmented FDR controlling procedure suggested in Van Der Laan et al. (2004). This procedure takes the joint density of the individual gene p-values in the gene set into account, whereas the BH procedure relies only on the marginal distribution of the p-values. Dudoit et al. (2004) suggest that there is substantial power gain of joint procedures compared to marginal procedures when the number of hypotheses tested is fairly small. In our gene sets, the median number of genes per gene set is 8 and the average is 30. Therefore, the minP-augmented FDR is expected to be much more powerful than the BH procedure within discovered gene sets. We refer to these two procedures as *screening+BH* and *screening+augmented minP*, and collectively as *screening+FDR*.

5.2 Results

Procedure 3.1 controls the OFDR under Microarray Dependency. The *estimated OFDR* is the average of the proportion of falsely discovered gene sets out of all gene sets discovered. It was below the nominal level of 0.05 for the screening+minP in all the settings considered (See column 2 of Table 4). Moreover, since OFDR control implies FDR control on the screening hypotheses, the simulations suggest that the FDR on the screening hypotheses is controlled under microarray dependency.

Table 4. The mean (SE) in 40 simulations of the total number of true gene set discoveries as well as the total number of individual gene discoveries in the screening+minP procedure, and the total number of individual gene discoveries in the gene be gene analysis.

Δ	Estimated OFDR	# True sets discovered	# True unique genes discovered	
			screening+minP	gene-by-gene
0.1	0.016 (0.005)	86.3 (19.7)	3.44 (1.82)	8.23 (0.64)
0.2	0.008 (0.003)	366.9 (26.9)	36.6 (0.5)	9.9 (0.4)
0.3	0.016 (0.004)	752.2 (25.5)	121.3 (0.65)	87.21 (2.58)
0.4	0.011 (0.003)	896.3 (16.5)	184.2 (0.4)	155.5 (0.4)

Table 5. The mean (SE) in 40 simulations of the total number of true gene set discoveries in a GSEA analysis as well as the total number of individual gene discoveries in the leading-edge subset analysis of GSEA.

Δ	Estimated OFDR	Estimated FDR of gene sets	Estimated FDR of genes	# True sets discovered	# True unique genes discovered
0.1	0.05 (0.03)	0.000 (0.003)	0.024 (0.107)	4.9 (21.4)	9.0 (40.1)
0.2	1 (0)	0.056 (0.022)	0.35 (0.02)	108.0 (21.9)	217.7 (17.9)
0.3	1 (0)	0.038 (0.012)	0.51 (0.04)	294.4 (26.7)	276.7 (6.6)
0.4	1 (0)	0.024 (0.015)	0.55 (0.08)	383.3 (14.6)	302.4 (2.0)

Procedure 3.1 may be more powerful than a gene-by-gene analysis for discovering individual genes. Table 4 shows the average of the number of true gene set discoveries as well as the total number of true gene discoveries for the screening+minP procedure, and the total number of gene discoveries for the gene-by-gene analysis. When the signal difference is weak $\Delta = 0.1$ barely any individual genes are discovered by either method, but 86.3 gene sets are discovered on average using the screening+minP procedure. As the signal increases more individual genes are discovered, and the screening+minP procedure has greater power than the gene-by-gene analysis.

Procedure 3.1 has advantages over GSEA. Table 5 shows that the average number of true gene set discoveries is much smaller for GSEA compared to Procedure 3.1. Moreover, it shows that while GSEA controls the FDR at the gene set level, the leading-edge subset analysis produces as many true individual gene discoveries as it does false gene discoveries. Therefore, even though this analysis discovers more than twice as many true gene discoveries as 3.1, the false positive rate is about 0.5. Since the leading-edge subset analysis always contains false positives, the estimated OFDR of GSEA is 1 for $\Delta \geq 0.2$. For $\Delta = 0.1$ the OFDR is small because gene set discoveries where made only in 2 out of the 40 runs. However, in the 2 runs that produced gene set discoveries about half of the genes discovered by a leading-edge subset analysis were false.

Screening+FDR procedures may be more powerful than Procedure 3.1 but they may not achieve OFDR control. Table 6 shows for every Δ the estimated OFDR for the screening+augmented minP and screening+BH procedures, as well as the average number of true individual gene discoveries. The screening+augmented minP is the most powerful procedure considered. For $\Delta = 0.2$ it discovers almost twice as many genes as the screening+minP procedure, but for $\Delta = 0.3$ it discovers only about 20% more genes and for $\Delta = 0.4$ only about 10% more genes. The screening+BH procedure appears to have very similar power to that of the screening+minP procedure. Note that there is no reason for either procedure to control the OFDR, even though the OFDR is controlled in this particular setting for the screening+BH procedure as well as for $\Delta > 0.1$ for the screening+augmented minP procedure.

Table 6. The estimated OFDR (and SE) and average of the total number (and SE) of true gene discoveries in 40 simulations of the screening+augmented minP and screening+BH procedures.

Δ	estimated OFDR		# True genes discovered	
	screening + augmented minP	screening + BH	screening + augmented minP	screening + BH
0.1	0.124 (0.019)	0.015 (0.004)	13.26 (0.33)	2.62 (0.16)
0.2	0.018 (0.006)	0.007 (0.004)	66.5 (0.7)	32.8 (0.7)
0.3	0.026 (0.005)	0.03 (0.008)	147.0 (0.5)	128.8 (0.7)
0.4	0.016 (0.003)	0.022 (0.004)	203.1 (0.3)	194.7 (0.4)

6 DISCUSSION

We have described a method, screening+minP, for testing multiple hypotheses on multiple gene sets. We introduced an appropriate error measure, the OFDR, as well as a general procedure for controlling this error measure.

We illustrated the usefulness of our approach in focusing attention on discovered gene sets where a large proportion of individual genes have been discovered. Our procedure controlled the proportion of falsely discovered screening hypotheses as well as falsely discovered gene sets. Thus, by applying the screening+minP procedure at level q and displaying the discovered gene sets and the discovered genes within a gene set in a table similar to Table 3, we expect at most q of the rows to contain false positives.

A comparison of the screening+minP method with the widely used GSEA method shows that GSEA may be less powerful in detecting differentially expressed gene sets. Moreover, the leading-edge subset analysis of GSEA may report many false positives.

The advantage of screening gene sets over a gene-by-gene analysis is large when the signal is weak since there may be enough power to detect differentially expressed gene sets even when there is practically no power to detect differentially expressed genes. Depending on the signal configuration, the screening+minP procedure may be more powerful in detecting individual genes than a gene-by-gene analysis. There does not appear to be a disadvantage in power of the screening+minP procedure over the screening+BH procedure. However, there is a disadvantage in power in comparison to the screening+augmented minP procedure when the signal is weak.

An essential feature of analysis of large data sets is to control for false positives in order to guarantee that the results are reproducible. We established theoretically that the screening+minP procedure controls the OFDR for independent p-values. We applied the method to the dependent situation of microarray analysis, based on evidence from simulations that the screening+minP controls the OFDR in such settings. Further verification of the validity of the screening+minP procedure in dependent settings is a direction for future research.

7 APPENDIX

7.1 Proof of Theorem 3.1

Let I_0 be the index set for gene sets that have at least one true null hypothesis. Let $V(s) = 1$ if gene set s is a falsely discovered gene set, i.e. a set that has been discovered and at least one of $H_{\text{screen}}(s), H_1(s), \ldots, H_{n(s)}(s)$ has been falsely rejected, and 0 otherwise. Let $R(s) = 1$ if gene set s is a discovered set and 0 otherwise. Let $Q = \frac{\sum_{s=1}^{S} V(s)}{\sum_{s=1}^{S} R(s)}$ if at least one gene set was discovered and 0 otherwise. Q is the proportion of falsely discovered gene sets out of all discovered gene sets if at least one gene set was discovered. The OFDR is, by definition, $E[Q]$. Then

$$E(Q) = \sum_{s=1}^{S}\sum_{k=1}^{S}\frac{1}{k}Pr(V(s)=1 \cap \sum_{i=1}^{S} R(i)=k)$$

$$= \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}Pr(V(s)=1 \cap \sum_{i=1, i\neq s}^{S} R(i)=k-1)$$

$$= \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}Pr(V(s)=1 \cap C_k^{(s)})$$

where $C_k^{(s)}$ is the event that $p_{(k-1)}^{(s)} \leq kq/S, p_{(k)}^{(s)} > (k+1)q/S, \ldots, p_{(S-1)}^{(s)} > q$, where $p_{(1)}^{(s)} \leq \ldots \leq p_{(S-1)}^{(s)}$ are the ordered coordinates of the vector $\vec{p}^{(s)}$ of p-values for the screening hypothesis excluding that of s.

$$E(Q) = \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}Pr(V(s)=1 \cap C_k^{(s)})$$

$$= \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}Pr(P_{\text{screen}}(s) \leq kq/S \cap \{\text{type I error at } s\} \cap C_k^{(s)})$$

$$= \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}Pr(P_{\text{screen}}(s) \leq kq/S \cap \{\text{type I error at } s\})Pr(C_k^{(s)})$$

where the last equality follows since the p-values are independent.

For every gene set with $R(s) = 1, s \in I_0$, there is a hypothesis that was falsely discovered. If the falsely discovered hypothesis is the screening hypothesis (i.e. the screening hypothesis is null), then $Pr(P_{\text{screen}}(s) \leq kq/S) \leq kq/S$. If the falsely discovered hypothesis is not the screening hypothesis, then the probability of rejecting such a null is $\leq kq/S$ since the FWER is controlled at level kq/S . Therefore $Pr(P_{\text{screen}}(s) \leq kq/S \cap \{\text{type I error at } s\}) \leq kq/S$ for a given k.

The result follows:

$$E(Q) \leq \sum_{s\in I_0}\sum_{k=1}^{S}\frac{1}{k}\cdot kq/S \cdot Pr(C_k^{(s)}) = \sum_{s\in I_0} q/S \leq q$$

REFERENCES

Bauer, P., Rohmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17:2133–2146.

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, doi: 10.1111/j.1541-0420.2007.00983.x.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93 (3):491–507.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics*, 171:783–790.

Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4):1165–1188.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (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.

Consortium, T. G. O. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Dudoit, S., Van Der Laan, M., , and Birkner, M. (2004). Multiple testing procedures for controlling tail probability error rates. *Technical report, University of California, Berkeley*, 166.

Ge, Y., Dudoit, S., and Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

Goeman, J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8).

Goeman, J., van de Geer, S., de Kort, F., and van Houwelingen, H. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Liu, J., Hughes-Oliver, J., and Menius, A. (2007). Domain-enhanced analysis of microarray data using go annotations. *Bioinformatics*, 23:1225–1234.

Nam, D. and Kim, S. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9 (3):189–197.

Nettleton, D., Recknor, J., and Reecy, J. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, 24(2).

Pollard, K., Ge, Y., Taylor, S., and Dudoit, S. (2008). Resampling-based multiple hypothesis testing. r bioconductor package version 1.20.0. *http://www.bioconductor.org/packages/2.1/bioc/html/multtest.html*.

Reiner, A. (2007). Fdr control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal*, 49(1):107–126.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375.

Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., and Mesirov, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences of the USA*, 102(43):15545–15550.

Tian, L., Greenberg, S., Won Kong, S., Altschuler, J., Kohane, I., and Park, P. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the national academy of sciences of the USA*, 102 (38):13544–13549.

Van Der Laan, M., Dudoit, S., and Pollard, K. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3 (1).

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

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103 (481):309–316.