

A Statistical Framework for Testing Functional Categories in Microarray Data

William T. Barry¹, Andrew B. Nobel², and Fred A. Wright³

1) Department of Biostatistics and Bioinformatics,
Duke University Medical Center, Durham, North Carolina 27710, U.S.A

2) Department of Statistics and Operations Research,
University of North Carolina at Chapel Hill, North Carolina 27599-3260, U.S.A

3) Department of Biostatistics,
University of North Carolina at Chapel Hill, North Carolina 27599-7420, U.S.A

Abstract

Many DNA microarray studies have shifted focus toward identifying the differential expression of groups of genes with shared biological function. We define a general framework for existing methods as contrasting a category to that of its complementary set of genes on the array. This includes tests for overrepresentation within a list of significant associations, and methods that consider quantitative measures of differential expression. Existing tests are divided into two classes. Class 1 tests assume gene-specific measures of differential expression are independent, despite overwhelming evidence of positive correlation. Analytic and simulated results demonstrate Class 1 tests are strongly anti-conservative in practice. Class 2 tests use array permutation to account for correlation, and by construction have proper Type I error control for the induced null. However, both classes use a null hypothesis that all genes have the same degree of differential expression. We introduce a more sensible and general (Class 3) null that the profile of differential expression is the same within the category and complement. Under this broader null Class 2 tests are shown to be conservative. We present a novel bootstrap test for the Class 3 null and demonstrate it provides valid Type I error control and more power in simulated and real microarray datasets.

1 Introduction

DNA microarrays allow researchers to simultaneously measure the coexpression of thousands of genes. They are widely used in biology and medicine to study the relationships between transcriptional expression and cellular processes or disease states. A primary application of microarrays is the identification of genes with differing expression across experimental conditions or with association to a clinical outcome. Hereafter we will generically refer to the condition or clinical outcome as the *response* for each array, and the association between expression and response as differential expression (DE).

Analyses of DE often proceed in a gene-by-gene manner, with the response compared to expression of each gene individually. The association for each gene is assessed using an appropriate statistic, and variety of methods have been proposed, including standard parametric tests, permutation-based methods, and Bayesian techniques (Dudoit et al. 2002; Tusher et al. 2001; Newton et al. 2004). These methods produce a ranked list of significant genes, with control of the family-wise error rate (FWER) or false discovery rate (FDR).

Although gene-specific analyses have yielded tremendous insight into the role of individual genes, they do not provide an organized framework for identifying larger-scale biological phenomena. With the ready availability of comprehensive annotation databases, such as Gene Ontology (GO) (Ashburner et al. 2000), researchers can now explore the coordinated involvement of *gene categories*, *i.e.*, sets of genes with shared annotation or function. A framework is warranted for examining the associations observed across an entire category to provide a more systematic understanding of DE.

Beginning with Virtaneva et al. (2001), a number of procedures have been pro-

posed to assess the association between a response and gene categories. The most commonly used tests begin with a list of genes deemed significant according to some criterion. A secondary analysis looks for over-representation, or *enrichment*, of genes within the category on the gene-list, using Fisher's Exact Test or other tests of 2 x 2 contingency tables (see Barry et al. (2005) for a list of references). Other approaches use the gene-specific measures of DE for all genes on the array. In these methods, category tests are constructed to compare a category to its complementary set of genes using an average difference of gene-specific statistics (Kim and Volsky 2005; Boorsma et al. 2005) or rank procedures for two-sample comparisons (Barry et al. 2005; Mootha et al. 2003; Ben-shaul et al. 2005).

Gene category testing is now widely performed, and results are frequently reported without independent verification. As pointed out in a recent review by Allison et al. (2006), even fundamental issues such as a formal definition of the underlying null hypothesis and a proper demonstration of Type I error have not been provided for the methods in the literature. There is thus a clear need to place gene category testing on a firm statistical foundation.

1.1 Contributions

In this paper we provide a careful examination of gene category testing by first defining a general framework which includes existing methods. Two distinct classes of procedures emerge with different null hypotheses:

1. Gene-specific statistics are independent and identically distributed; or
2. Gene-specific statistics follow a common null distribution, though may be dependent.

Several shortcomings of the two classes of procedures are revealed through analytic derivation and simulations from an example dataset.

We then propose a broader null hypothesis that allows for varying degrees of association between gene-specific expression and response, and also allows for dependence of expression between genes. Under this general category null, array permutation approaches can be quite conservative. The conservativeness can be explained in part through an analytical argument which shows that the maximum variance of the category-wide test statistic occurs under the special case induced by array permutation. We present a simple and powerful bootstrap-based approach that is consistent with the more general null hypothesis. Finally, we demonstrate the utility of this new method in a breast cancer dataset, and discuss other advantages that bootstrap-based tests have over array permutation procedures.

2 Notation and general framework for gene category tests

Let \mathbf{x} be an $m \times n$ matrix containing the observed expression data for an experiment with m genes and n arrays; let x_{ij} be the element of the matrix corresponding to the i -th gene in the j -th array. The expression profile for gene i is the row vector \mathbf{x}_{i*} , and the expression values of array j are represented by the column vector \mathbf{x}_{*j} . We use lowercase to denote observed values, and uppercase (*i.e.*, \mathbf{X} , X_{ij} , \mathbf{X}_{i*} and \mathbf{X}_{*j}) to denote random versions of these quantities. The array-specific response information is denoted by \mathbf{y} , with element y_j corresponding to array j . The response can be categorical (*e.g.*, tumor grade or experimental group assignment) or continuous (*e.g.*, survival time), and could potentially be multivariate. A category is represented by a subset $C \subseteq \{1, \dots, m\}$ such that $i \in C$ if and only if gene i is a member of the category. The size of a category C will be denoted by $m_C = \sum_{i=1}^m I\{i \in C\}$. For any

category C , the complementary set of genes will be denoted by \bar{C} and will be of size $m_{\bar{C}} = m - m_C$.

We adopt the terminology of Barry et al. (2005), who pointed out that hypothesis tests of gene categories can be viewed as a two-stage procedure (See Box 1). In the first stage, a *local statistic* measures the association between the expression profile of each gene and the response. We denote the local statistic of gene i by $T_i = T(\mathbf{X}_{i*}, \mathbf{y})$ and let t_i be the corresponding value from observed data. In a two-condition experiment, the local statistic might be a t -statistic or average fold change, while in more complex experimental designs, such as censored survival data, a local statistic derived from the Cox proportional hazard model may be used to test for an association between gene expression and patient outcome. For many common experiments, T reflects an underlying gene-specific parameter of association between response and expression. In the two-condition example, the related parameters would be a scaled mean difference and a ratio of population means, respectively. Properties of local statistics are examined more fully in Section 5.3.

In the second stage of a gene category test, a *global statistic* is used to compare the local statistics of genes within a category C to those in its complement. We denote the global statistic by $U = U(T_1, \dots, T_m : C)$, and in the following sections describe the functional forms of $U(\cdot)$ that have already been proposed in the literature. Existing methods focus on either detecting a difference in the proportion of genes called significant, or detecting a shift in the average local statistic within the category versus its complement.

There are various ways to classify existing gene category tests, e.g. by the choice of global statistic, or whether permutation of genes or arrays is used. In terms of Type I error control, we argue that the most meaningful distinction is whether or not *array* permutation is used. This distinction is used to divide existing procedures into

two classes, and it is further instructive to re-state the classes in terms of the null hypotheses for which the procedures have appropriate error control.

Box 1: Common elements of gene category tests

Gene category tests are typically two-stage procedures requiring the following statistics:

- A *local statistic* that measures the association between response (*e.g.* experimental condition) and expression of each gene.
- A *global statistic* that compares the local statistics within a category to those of its complement.

Two classes of hypothesis tests are typically designed for each global statistic:

1. Parametric or rank-based procedures that assume independent and identically distributed local statistics, or gene permutation methods that induce essentially the same null.
2. Array permutation methods which induce a null that maintains the correlation structure among genes while removing all associations with the response.

Error rate controlling or estimating procedures address the multiple comparisons involved in simultaneously testing a number of different gene categories.

3 Class 1 gene category tests

Gene category test statistics are intended to be sensitive to an increase in the DE of genes within a category compared to the genes in its complement. For many gene category test procedures, hypothesis tests are performed using traditional methods for comparing independent samples from two populations. We note that for these

methods, the null hypothesis is rarely stated, and it is rarely discussed whether the independence assumptions are met. These tests vary in terms of the global statistic employed and whether exact or approximate distributions are used to determine p -values, but share the following common null hypothesis.

Definition 1 *Class 1 gene category tests are defined by the assumed or induced null hypothesis that the local statistics T_1, \dots, T_m , are independent and identically distributed (i.i.d.). More precisely,*

$$H_0 : T_1, T_2, \dots, T_m \text{ are i.i.d with } T_i \sim F \quad (1)$$

where F can take any form.

3.1 A survey of the global test statistics

The global statistics that have been proposed for Class 1 tests can be identified as “categorical” when a list of significant genes has been previously identified by a gene-specific analysis, or “continuous” when a more direct measure of DE is used for each gene. To illustrate the variety of global statistics that have been proposed, we present two examples of each case and a brief description of the corresponding nonresampling-based Class 1 tests. A one-sided form of each test is given, because in most applications one is only interested in categories showing increased association with the response in the category as compared to its complement.

Categorical statistics. Gene-list enrichment methods have developed as a *post hoc* means of testing a category once genes with significant DE have been identified. Let Γ denote the (possibly data-dependent) rejection region for the local statistics that produces the list of significant genes. These methods consider only the dichotomous outcomes of the m gene-specific hypothesis tests, $I\{T_i \in \Gamma\}$, and the DE within C and \bar{C} is therefore summarized by a 2×2 contingency table (Figure 1).

The traditional contingency table tests that have been proposed for gene category analysis include the χ^2 test of homogeneity, Fisher’s Exact test, and slight variations on these. In the classical derivation of these tests, the binary variables $I\{T_1 \in \Gamma\}, \dots, I\{T_m \in \Gamma\}$ are assumed to be independent with the probabilities of rejection $P(T_i \in \Gamma) = \pi_C$ for $i \in C$ and $P(T_i \in \Gamma) = \pi_{\bar{C}}$ for $i \in \bar{C}$. The tests look for departures from $\pi_{\bar{C}} = \pi_C$, under the assumption that the indicator variables are *i.i.d.* It is worthwhile to note that the Class 1 null in (1) is sufficient but not necessary for the dichotomous outcomes to be *i.i.d.* under a given Γ . However, (1) guarantees the categorical null holds for any possible choice of rejection region.

In several of the gene-list enrichment software packages, the χ^2 test of homogeneity is proposed as an approximate test for large categories (Draghici et al. 2003; Beißbarth and Speed 2004). The one-sided version of this test is equivalent to the difference in proportions test proposed originally by Pearson (1911), where the global statistic can be written as

$$U_P = \hat{\pi}_C - \hat{\pi}_{\bar{C}} = \frac{1}{m_C} \sum_{i \in C} I\{T_i \in \Gamma\} - \frac{1}{m_{\bar{C}}} \sum_{i' \in \bar{C}} I\{T_{i'} \in \Gamma\}. \quad (2)$$

By the central limit theorem, the two proportions are asymptotically normal for large m_C and $m_{\bar{C}}$, and thus a Z-test can be performed on a standardized form of U_P .

Fisher’s Exact Test is more commonly applied in gene-list enrichment methods. Formally it is a conditional test based on the total number of rejected hypotheses, $R = \sum_{i=1}^m I\{T_i \in \Gamma\}$. The global statistic can be represented as the number of genes in the category that are rejected, namely

$$U_F = \sum_{i \in C} I\{T_i \in \Gamma\}. \quad (3)$$

Under (1) an exact one-sided p -value can be obtained from the hypergeometric distribution. Depending on how the gene-list is obtained, it is not always clear that it is appropriate to condition on R , but exact tests are often favored in order to

handle small categories. For moderately sized categories, we note there will be little difference between the exact conditional and approximate unconditional tests.

Continuous statistics. In contrast to gene-list type tests, it is also possible to directly compare the observed associations of expression to response without an intermediate list. One straightforward global statistic is the average difference in local statistics between category and complement, namely

$$U_D = \frac{1}{m_C} \sum_{i \in C} T_i - \frac{1}{m_{\bar{C}}} \sum_{i' \in \bar{C}} T_{i'} . \quad (4)$$

Two related hypothesis tests based on U_D have been proposed in the literature. In one, a t -test is performed after standardizing U_D by the pooled sample variance of the local statistics (Boorsma et al. 2005). In another, a Z -test is performed after U_D is scaled by the standard deviation of all local statistics (Kim and Volsky 2005). For a typical category where $m_C \ll m$, the variance estimates in the two approaches will be similar, yielding comparable test results.

The global statistic in (4) may not be robust to skew or outlying observations in the local statistics. Rank-based global statistics avoid this shortcoming, as they are invariant to monotone transformations of the local statistics. The Wilcoxon rank sum test has been implemented in the software GOSTat (Beißbarth and Speed 2004). In the absence of ties the global statistic can be written as

$$U_W = \sum_{i \in C} \text{Rank}(T_i) . \quad (5)$$

Under the Class 1 null hypothesis the discrete CDF of U_W is known once m_C and $m_{\bar{C}}$ are specified. Hypothesis testing proceeds using an exact procedure or a normal approximation to U_W .

A Kolmogorov-Smirnov type global statistic has also been implemented in another rank-based Class 1 procedure (Ben-shaul et al. 2005). However, the Kolmogorov-Smirnov statistic has been criticized in gene category testing for being sensitive to

departures that do not necessarily reflect increasing DE in the category (Damian and Gorfine 2004). For example, a category with no DE but with local statistics that all happen to be very close to one another would be identified as significant by these tests. For this reason, we restrict our focus to U_D and U_W when considering continuous global statistics.

3.2 Gene Permutation

Several permutation-based methods have proposed randomly reordering the rows of the data matrix to determine category significance (Ashburner et al. 2000; Zhong et al. 2004; Pavlidis et al. 2004). In this setup, the collection of local statistics remains unchanged while the category assignments are randomized. This resampling scheme approximately induces the Class 1 null hypothesis in (1) with each reassigned local statistics following the empirical distribution of the observed values. $\hat{F}(t) = m^{-1} \sum_{i=1}^m I\{t_i \leq t\}$. Thus, although gene permutation may be useful, e.g. in producing calibrated versions of complicated global statistics, such procedures belong to Class 1. In fact, a Fisher’s Exact test of U_F and Wilcoxon rank sum test of U_W will be identical to the exhaustive permutations of gene assignment to a category.

4 The effect of correlation on Class 1 tests

In this section we examine more closely the assumption of independent local statistics, and its failure to hold in gene expression data. We relate the correlation in expression to that of local statistics, and show how it affects Class 1 tests. A simulation study based on real microarray data exhibits the extreme anti-conservative behavior of Class 1 tests in the presence of realistic levels of expression correlation.

4.1 Correlations in expression and local statistics

Let the population correlation between genes i and i' be given as $\rho_{i,i'}^X = \text{Corr}(X_{ij}, X_{i'j})$. For experimental designs with independent arrays, a natural estimate of $\rho_{i,i'}^X$ is the sample correlation coefficient

$$r_{i,i'} = \frac{\sum_{j=1}^n (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \cdot \sum_{j=1}^n (x_{i'j} - \bar{x}_{i'})^2}} \quad (6)$$

where $\bar{x}_i = n^{-1} \sum_{j=1}^n x_{ij}$.

The true distributions of global statistics for Class 1 tests are directly affected by the correlation between local statistics, $\rho_{i,i'}^T = \text{Corr}(T_i, T_{i'})$. In the special case that T takes the linear form $T(\mathbf{X}_{i*}, \mathbf{y}) = \sum_{j=1}^n a(y_j) \cdot X_{ij}$ for some function $a(\cdot)$, it is easy to show that $\rho_{i,i'}^T = \rho_{i,i'}^X$. An example of a linear local statistic would be an unscaled difference in sample means, *e.g.*, fold change on the log-scale.

In general, the relationship between the correlations $\rho_{i,i'}^X$ and $\rho_{i,i'}^T$ does not have a simple analytic form, although it can be shown numerically to be monotone and often nearly linear for one-sided local statistics. Monte Carlo simulations of gene expression data (Figure 2) demonstrate this relationship holds in several standard experimental designs and corresponding measures of DE, including t -statistics for two-condition studies and for regressing expression on censored time-to-event data through a Cox proportional hazards model. For such local statistics, $\rho_{i,i'}^X \approx \rho_{i,i'}^T$, so that the sample correlation coefficients can also estimate $\{\rho_{i,i'}^T\}$. The relationship may be nonlinear for “undirected” local statistics, such as an analysis of variance F-statistic (Figure 2B)

4.2 Variance inflation

The effects of pairwise correlation on Class 1 tests are demonstrated through deriving the true variances of the global statistics U_D and U_W . The distributions of the

categorical global statistics U_F and U_P are similarly effected by correlation, but are not presented because the variances will depend on both the underlying distribution of local statistics T and the rejection region Γ .

For the average difference global statistic U_D , the true variance will differ from that under the *i.i.d.* null in Class 1 tests by three additional terms

$$\text{Var}[U_D] = \text{Var}_{i.i.d.}[U_D] \left(1 + \frac{m_{\bar{C}}(m_C - 1)}{m} \rho_C + \frac{m_C(m_{\bar{C}} - 1)}{m} \rho_{\bar{C}} - \frac{m_C \cdot m_{\bar{C}}}{m} \rho_{C, \bar{C}} \right) \quad (7)$$

where

$$\rho_C = \frac{1}{m_C \cdot (m_C - 1)} \sum_{i \in C} \sum_{\substack{i' \in C \\ i' \neq i}} \rho_{i, i'}^T \quad (8)$$

$$\rho_{\bar{C}} = \frac{1}{m_{\bar{C}} \cdot (m_{\bar{C}} - 1)} \sum_{i \notin C} \sum_{\substack{i' \notin C \\ i' \neq i}} \rho_{i, i'}^T \quad (9)$$

$$\rho_{C, \bar{C}} = \frac{1}{m_C \cdot m_{\bar{C}}} \sum_{i \in C} \sum_{i' \notin C} \rho_{i, i'}^T. \quad (10)$$

relating to the average pairwise correlation within the category (8), within the complement (9), and across the two gene sets (10). We note that ρ_C can vary greatly across categories, while $\rho_{\bar{C}}$ and $\rho_{C, \bar{C}}$ will be close to the average correlation across the array and near zero in most datasets. In that case, for a moderately sized category where $m_{\bar{C}} \approx m$ the variance inflation over the Class 1 assumption, $\text{Var}[U_D]/\text{Var}_{i.i.d.}[U_D]$, will be approximated by $1 + (m_C - 1) \cdot \rho_C$. Thus, categories exhibiting positive correlation will have a U_D global statistic with greater variance than what is assumed under (1), leading to an anti-conservative Class 1 test.

For the Wilcoxon rank sum global statistic, the true variance will depend on the underlying distribution of local statistics, F , in (1). The following theorem derives $\text{Var}[U_W]$ for normally distributed local statistics.

Theorem 1 *Let T_1, \dots, T_m be identically distributed random variables that follow a multivariate normal distribution with unit variances and pairwise correlations $\{\rho_{i, i'}^T\}$.*

Then for a given category, $C \subset \{1, \dots, m\}$, the variance of $U_W = \sum_{i \in C} \text{Rank}(T_i)$ is

$$\text{Var}[U_W] = \frac{1}{2\pi} \sum_{i \in C} \sum_{i' \in C} \sum_{h \notin C} \sum_{h' \notin C} \sin^{-1} \left(\frac{\rho_{i,i'}^T + \rho_{h,h'}^T - \rho_{i',h}^T - \rho_{i,h'}^T}{\sqrt{(2 - 2\rho_{i,h}^T) \cdot (2 - 2\rho_{i',h'}^T)}} \right). \quad (11)$$

(see Appendix A for the proof). In the special case that the local statistics within a category are positively correlated, while all genes in the complement are independent, this variance can be shown through simple algebra to be strictly greater than the variance assumed under the Class 1 null. Despite these analytical solutions for special cases, in general the pairwise correlations, $\{\rho_{i,i'}^T\}$, are unknown. Thus, to better explore the consequences of correlation for categories from a real microarray dataset, we have applied the four Class 1 tests to a simulation study described below.

4.3 A simulation study

A two-condition experiment was simulated using a subset of the lung carcinoma microarrays from Bhattacharjee et al. (2001). 100 adenocarcinoma samples were arbitrarily selected with expression estimates for 7299 genes (see Barry et al. (2005) for data pre-processing steps); 1823 GO and Pfam categories were identified with at least 5 members among the expressed genes. The within-category average pairwise sample correlations ranged from -0.09 to 0.93, with more than 86% of the categories having values greater than the average pairwise correlation across the entire array (0.012). This increase in correlation within categories reflects the fact that coexpression among genes is related to function (Lee et al. 2004).

1000 response vectors were randomly generated to assign each array to one of two conditions with equal sample size. This ensured that the simulated datasets had no association between expression and experimental condition for any gene, and thus no category should have greater DE than any other. We note that the expression matrix was held constant across simulations, so the sample gene-gene correlations $\{r_{i,i'}\}$ remained fixed.

For each realization of the response vector, the absolute value of a pooled-variance t -statistic was used as the local statistic, and global statistics U_F , U_P , U_D , and U_W were calculated. For the Fisher’s Exact Test statistic, U_F , and the difference in proportions, U_P , the rejection region was set as values exceeding $t_{98,0.975} = 1.984$. For each global statistic and each category, Class 1 tests yielded a nominal p -value for every realized response vector. Histograms of the nominal p -values pooled across all categories and all realizations demonstrate their extreme non-uniformity under the induced null hypothesis, confirming the poor performance of Class 1 tests (Figure 3).

The average Type I error of these tests is estimated as the proportion of p -values under simulations that fall below a target α level. For each global statistic, the corresponding Class 1 test becomes more anti-conservative for smaller target α values (Table 1). While the gene-list enrichment methods are slightly less anti-conservative than the continuous methods, this is offset by their potential loss in power from dichotomizing local statistics.

To illustrate how this behavior also affects the family-wise error rate among the $L = 1823$ categories, we applied a Bonferroni correction to the nominal p -values. Since for the randomized data all categories are truly null, the FWER is estimated by

$$FWER = \frac{1}{1000} \sum_{b=1}^{1000} I \left\{ \sum_{h=1}^L I \left\{ p_{b,h} < \frac{\alpha}{L} \right\} > 0 \right\} \quad (12)$$

where $p_{b,h}$ is the Class 1 p -value for category h under realization b . There is substantial overlap in the membership of gene categories from annotations such as Gene Ontology. Therefore, the use of Bonferroni thresholds might be thought to be conservative in controlling the FWER, providing some protection against anti-conservative Class 1 p -values. However, for $\alpha = 0.05$, the realized FWER in (12) is far greater than the target level ($U_F : 0.776$, $U_P : 0.910$, $U_D : 0.925$, and $U_W : 0.918$). The extreme anti-conservativeness of the Class 1 tests of all four global statistics suggests a different

approach is needed to conduct valid gene category tests.

5 Class 2 tests and array permutation

5.1 Defining the Null Hypothesis

The null hypothesis of Class 1 tests is violated by the correlations present in gene expression data, and we have demonstrated the negative effect they have on Class 1 tests. For this reason, a second class of gene category tests is warranted that can identify increases in differential expression within a category, while accounting for correlation.

Definition 2 *Class 2 gene category tests are defined by the assumed or induced null hypothesis, that the local statistics T_1, \dots, T_m , are identically distributed while possibly dependent. More precisely,*

$$H_0 : T_1, T_2, \dots, T_m \text{ are identically distributed with } T_i \sim F_0 \quad (13)$$

where F_0 corresponds to a lack of association between expression and the response of interest.

The Class 2 null hypothesis in (13) is stated in terms of the local statistics being identically distributed under the null hypothesis. To more fully describe a sufficient property of local statistics to ensure (13), we revisit the process of selecting an appropriate form of $T(\cdot)$ for a certain experimental design.

5.2 δ -determined local statistics

When testing for increased DE within gene categories, investigators are generally interested in a particular gene-specific association with the response. For many common

experimental designs, an unknown gene-specific parameter, δ_i , meaningfully captures this association. In order to conduct a gene-specific analysis of DE, $T(\cdot)$ is chosen as a measure that can be used in hypothesis tests against a null value for the parameters of every gene $\{\delta_i\}$. As illustration, consider a two-condition experiment where the response vector \mathbf{y} takes values of 1 or 2, indicating the sample condition of each array. If the expression of gene i has expectation μ_{1i} and μ_{2i} under the two conditions, and common variance σ_i^2 , then the underlying association of interest in these experiments can be represented as the scaled difference in means

$$\delta_i = \frac{\mu_{1i} - \mu_{2i}}{\sigma_i \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} . \quad (14)$$

In this case, the gene-specific hypothesis of interest is $H_{0,i} : \delta_i = 0$ and the pooled-variance t -statistic is a natural choice of local statistic (Galitski et al. 1999). When expression is normally distributed, the local statistic follows a central t -distribution when $\delta_i = 0$ and more generally a t -distribution with noncentrality parameter, δ_i .

In general, a function $T(\cdot)$ is a proper choice of test statistic for a null of the form, $H_{0,i} : \delta_i = d$, when the distribution $F(T_i | \delta_i = d)$ is known and does not depend on any nuisance parameters. When the distribution of $T(\cdot)$ can be specified in this manner for any choice of d , we refer to it as δ -determined. This property is important in the theory of interval estimation and pivotal quantities. If $F(T_i | \delta_i = d)$ is δ -determined, it can be used as a pivotal quantity to construct a confidence set for δ_i (Casella and Berger 2002). In the particular example presented above, a Student's t is δ -determined by (14).

The δ -determined property is also important when conducting gene category tests, because differences in nuisance parameters do not influence the comparison of a category against its complement. We illustrate the ramifications with the two-condition experiment and δ as defined in (14). Here the gene-specific means and

variances of expression are considered nuisance parameters. Suppose that for each gene one directly uses the modified t -statistic from the SAM software (Tusher et al. 2001) as the local statistic. This statistic contains a constant in the denominator that effectively penalizes lowly-expressed genes in order to improve the FDR for lists of rejected genes. The SAM t -statistic is not δ -determined, because its distribution will further depend on the means and variances of expression. Consider a category consisting of mainly highly-expressed genes (e.g., “housekeeping” genes). Even if no genes were differentially expressed across conditions, and thus no category should be considered special in this regard, the highly expressed genes in the category have an increased chance to appear amongst the most-significant genes in a ranked list. The category would thus falsely appear to be significant. Categories with lowly-expressed genes would experience the opposite effect, and would be less likely to be considered significant.

When δ -determined statistics are chosen, the Class 2 null can be restated in terms of the gene-specific parameters for all genes, $H_0 : \delta_1 = \dots = \delta_m = d_0$, where d_0 relates to there being no association between expression and response. For the remainder of the paper, we will only consider local statistics that are δ -determined, or approximately so when n is large.

5.3 Array Permutation

If the pairwise correlations in local statistics, $\{\rho_{i,i'}^T\}$, were known, a Class 2 test can be constructed for the average difference statistic, U_D , using its true variance derived in (7). Similarly, an approximate Z-test for the Wilcoxon rank sum statistic, U_W , can be designed using (11) if local statistics are approximately normal. However, since the correlations are generally unknown, a particular form of permutation can be used as an alternative means of approximately inducing the Class 2 null.

In many common microarray experiments, each mRNA sample constitutes an independent unit. By permuting the column vectors of \mathbf{X} , or equivalently the response vector, \mathbf{y} , an empirical null distribution is achieved in which there is no association between gene expression and the response. Array permutation was first used in Virtaneva et al. (2001) to test categories of genes, and then implemented in GSEA for a Kolmogorov-Smirnov global statistic (Mootha et al. 2003), and in SAFE for a Wilcoxon rank sum global statistic (Barry et al. 2005). More recently, other global statistics have been proposed including a weighted version of GSEA (Subramanian et al. 2005) and a standardized truncated mean that would be more sensitive to directional changes (Efron and Tibshirani in press). As noted in these publications, array permutation does not change the correlations in expression among genes, and thus local statistics will remain correlated in the resampled data. Since the resampled local statistics are conditional on the observed dataset, their empirical distributions are not likely to be exactly identically distributed, and only approximately follow the Class 2 null. However, if one converts to using the induced empirical p -values as local statistics, every gene will exactly follow the discrete uniform distribution under permutation, guaranteeing (13).

5.4 Simulated coverage of Class 2 tests

The simulation based on the Bhattacharjee et al. (2001) dataset was used to evaluate the Class 2 tests of each global statistic based on array permutation. Here the tests are ensured to be of proper size, since both the randomization procedure in the simulation and array permutation employ the same sampling schemes. We confirmed this by obtaining empirical p -values for each category and each realization of the response vector (Table 1). Due to computational restrictions the minimum possible empirical p -value was 0.001. The Class 2 Fisher’s Exact Test results are notably conservative,

due to the numerous tied global statistics that occur in small categories, but the slight misspecification of Type I error for U_P , U_D and U_W reflects only sampling variability.

These results demonstrate that Class 2 tests of gene categories clearly outperform Class 1 tests by accounting for the positive correlation in gene expression data.

6 A more general null for gene category tests

Although permutation-based Class 2 procedures properly acknowledge the array as the sampling unit, we note that they share with Class 1 and other Class 2 procedures the shortcoming of assuming a null hypothesis under which the marginal distribution of local statistics are identically distributed. This assumption is not necessary in establishing whether or not a relative increase in the amount of differential expression is observed within a category. For example, if 20% of the genes on the array are differentially expressed to the same degree, and the remaining genes are not differentially expressed, any category with 20% DE genes should not be considered “special.” However, the array permutation null is violated under this scenario.

Based on this simple example, we propose the following less restrictive and more biologically sensible null hypothesis. Instead of requiring all local statistics to be identically distributed, we allow the statistics of each gene to fall into one of $K \leq m_C$ strata, where each stratum corresponds to a different marginal distribution. The equality of a category of genes relative to its complement can then be formally stated as follows.

Definition 3 *Let $K \geq 1$ and distributions G_1, \dots, G_K be fixed. For a gene category, $C \subset \{1, \dots, m\}$, let the local statistics, T_1, \dots, T_m , be marginally distributed as F_1, \dots, F_m . Assume that each $F_i \in \{G_1, \dots, G_K\}$ and let $\beta_{C,k} = m_C^{-1} \sum_{i \in C} I\{F_i =$*

$G_k\}$ and $\beta_{\bar{C},k} = m_{\bar{C}}^{-1} \sum_{i \in \bar{C}} I\{F_i = G_k\}$ be the proportions of genes from the category, C , and complement, \bar{C} , respectively, whose local statistics are distributed as G_k . The Class 3 null hypothesis can be stated as follows.

$$H_0 : \beta_{C,k} = \beta_{\bar{C},k} \equiv \beta_k \quad k = 1, \dots, K \quad (15)$$

where the distributions, G_1, \dots, G_K can take any form.

Under this definition of the Class 3 null, we note that the Class 1 and Class 2 null hypotheses are special cases of (15), with $K = 1$ stratum. For experimental designs where DE can be expressed in terms of a δ and local statistics are chosen to be δ -determined, *e.g.*, (14) and the Student's t -statistic, then the strata of the Class 3 null can be related to the different degrees of association each gene has with the response. In this case, another way of stating (15) is that the empirical distribution of gene-specific parameters is identical between the two sets.

In the following subsections we will describe simple bootstrap-based tests that maintain the correlation structure of the expression data, and have approximately correct Type I error under the Class 3 null. The distributional properties of U_W are derived under (15), providing insight into why the bootstrap procedure has improved power over Class 2 procedures. Our earlier simulation study is adapted to recreate the Class 3 null, enabling us to quantify the improved error control of a Class 3 test in realistic data. The simulations will also demonstrate increased power of a Class 3 test under defined alternative hypotheses.

6.1 Defining the bootstrap-based tests

Standard bootstrap methodology assumes that the observed data can be divided into independent units derived from an unknown probability model. Resampling from the empirical distribution of the observed data enables one to form approximate confidence intervals without parametric assumptions (Efron and Tibshirani 1998).

For most microarray experiments, the independent sampling unit is the joint vector $\{\mathbf{x}_{*j}, y_j\}$ containing both the m gene expression measurements and response information for an mRNA sample. To approximate the unknown probability model of the data, we resample the joint vectors with replacement. Let $\mathbf{b} = (b_1, \dots, b_n)$ be a resampling vector whose elements are independent and uniformly distributed over the integers $\{1, \dots, n\}$. Associated with \mathbf{b} is a resampled response $\mathbf{y}^{*b} = (y_{b_1}, \dots, y_{b_n})$, and a resampled expression matrix in which the measurements of gene i are given by $\mathbf{x}_i^{*b} = (x_{ib_1}, \dots, x_{ib_n})$. From the resampled data, local statistics $t_i^{*b} = T(\mathbf{x}_i^{*b}, \mathbf{y}^{*b})$, and a global statistic $u^{*b} = U(t_1^{*b}, \dots, t_m^{*b} : C)$ may be calculated in the usual way. Let B denote the total number of bootstrap samples.

We use standard methods to generate bootstrap confidence intervals for the parameter $\theta = E[U]$, where the global statistic, U , is suitably chosen so that the expectation is known under the stratified null H_0 in (15). The corresponding hypothesis test determines if $\theta_0 = E_{H_0}[U]$ falls in the constructed interval. The following theorem establishes the expectation of the Wilcoxon global statistic U_W under all realizations of the stratified null hypothesis (15)

Theorem 2 *Suppose that the local statistic of each gene, T_i , has a distribution, $F_i \in \{G_1, \dots, G_K\}$. For any category, $C \subset \{1, \dots, m\}$, where for $\beta_{C,k} = \beta_{\bar{C},k} = \beta_k$ for every stratum of genes, $k = 1, \dots, K$ and where $P(T_i = T_j) = 0$ for $i \neq j$, then the expectation of U_W is*

$$E[U_W] = \frac{m_C \cdot (m + 1)}{2} . \quad (16)$$

(see Appendix B for proof). This expectation is constant, regardless of the number of strata K , the proportion of genes in each $\{\beta_1, \dots, \beta_K\}$, and distributions $\{G_1, \dots, G_K\}$. Similar derivations demonstrate that the global statistics U_D and U_P have a fixed expectation of 0 under (15). By contrast, the expectation of the global statistic

employed in Fisher’s Exact test depends on the K gene-specific distributions, and the expectation of the Kolmogorov-Smirnov type global statistic used in Mootha et al. (2003) depends on both the marginal distribution and correlation structure among local statistics. Thus, standard bootstrapped confidence intervals can not be used to conduct hypothesis tests for these global statistics. We favor the Wilcoxon rank sum, U_W , as it is a robust statistic that avoids the arbitrariness of choosing a rejection region for the gene-list methods; U_W is used as the global statistic for the remaining sections of the paper.

In order to test the null in (15) against one-sided alternatives of increased amounts of DE in the gene category, we produce a confidence interval for U_W with a lower bound L_α , the α -th percentile of U_W . The associated test rejects H_0 when $\theta_0 < L_\alpha$. A basic procedure for producing a confidence interval via bootstrap resampling is the quantile method (Efron 1979) where L_α is simply the sample α -quantile of the resampled values: $u_{(B-\alpha)}^*$. The quantile method is straightforward to compute and invariant under monotone transformations of the global statistics. However, its error control may be poor, especially when the sample size is small (Efron 1987) due to the difficulty of estimating the tail distribution of the global statistic.

Alternatively, if one assumes that the global statistic is approximately normal, a confidence interval can be generated from the t -distribution using bootstrap-based estimates of the moments of U_W (Efron 1979). The resulting one-sided confidence interval has a lower bound given by

$$\bar{u}^* - \hat{se}^*(U) \cdot t_{n-1, 1-\alpha} \tag{17}$$

where

$$\bar{u}^* = \frac{1}{B} \sum_{b=1}^B u_b^* \quad \text{and} \quad \hat{se}^*(U) = \left[\frac{\sum_{b=1}^B (u_b^* - \bar{u}^*)^2}{B-1} \right]^{\frac{1}{2}} \tag{18}$$

From (5) we note that U_W is the sum of $m_C \cdot (m - m_C)$ pairwise comparisons of

local statistics. When the average correlation between these terms is not extreme and m_C is large, approximate normality of U_W follows from the Central Limit Theorem. Histograms of resampled global statistics confirm that the approximation to the normal distribution is appropriate for the large number of genes in a typical microarray experiment. One advantage of the t -interval method over the quantile interval method is that the maximum attainable significance level is not bounded by the number of resamples B . Our simulations suggest that $B = 200$ arrays are typically sufficient for estimating the first two moments.

6.2 Type I error under a simulated null

The randomized lung cancer dataset described in Section 4.3 was used to evaluate the Type I error incurred by permutation- and bootstrap-based tests of U_W under the stratified null hypothesis. In the two-condition experiment with the absolute value of a Student's t as the local statistic, the stratum of distributions can be stated in terms of the δ given in (14). Several null hypotheses were investigated with $K = 2$ classes of genes relating to no DE ($\delta_i = 0$) and positive DE ($\delta_i = d > 0$). To artificially generate different degrees of DE in a particular gene, the expression values were first standardized to have variance 1; then $d \cdot \sqrt{1/n_1 + 1/n_2}$ was added to x_{ij} where $y_j = 1$. Simulations were run using three different levels of DE, $d = 1, 3$ and 5 , and also for three proportions of DE, $\beta = 1/5, 1/3$ and $1/2$. For each proportion, a subset of non-overlapping categories was selected such that for each, $\beta \cdot m_C$ was an integer. This resulted in 41 categories being considered for $\beta = 1/5$, 40 categories for $\beta = 1/3$, and 34 categories for $\beta = 1/2$. The selected categories exhibited a wide range of correlation in expression, reflective of that seen across the entire dataset.

For each of 1000 randomizations of tumor status, the Class 2 permutation- and Class 3 bootstrap-based hypothesis tests were conducted using 2500 permutations and

resamples, respectively. Type I error was determined by comparing the empirically derived p -values to various α levels (Figure 4). For a target $\alpha = 0.05$, the bootstrap Type I error was only slightly inflated, and remained relatively unchanged regardless of β and d , whereas the Type I error of permutation testing dropped dramatically as either parameter diverged from 0. For $d = 3$ and $\beta = 1/3$, the minimum empirical p -value obtained under permutation was 0.012 (Figure 4C). These findings illustrate the Class 2 test based on array permutation is overly conservative under the broader null.

The poor performance of permutation-based testing can be attributed to the fact, noted above, that a special case of (15) is induced under which the local statistics are approximately identically distributed (13). In order to better understand the conservative behavior, we return to the variance of the Wilcoxon global statistic derived in (11), and define the following type of positively correlated category

Definition 4 *For local statistics T_1, \dots, T_m with correlations $\{\rho_{i,i'}^T\}$, a category $C \subseteq \{1, \dots, m\}$ will be called **correlation dominant** if for every $\{i, i'\} \in C$ and $\{h, h'\} \notin C$ it is true that $\rho_{i,i'}^T \geq \rho_{i,h}^T$ and $\rho_{i,h}^T \leq \rho_{h,h'}^T$, so that all correlations within the category and within the complement are greater than those across the two gene sets.*

The correlation structure described in Section 4.2 is a particular example of a correlation dominant category. In the following theorem we establish that for normally distributed local statistics, the variance of the Wilcoxon global statistic U_W is maximized under the $K = 1$ null in (13) for correlation dominant categories.

Theorem 3 *Let T_1, \dots, T_m be random variables that follow a multivariate normal distribution with means $\delta_1, \dots, \delta_m$, unit variances and correlations $\{\rho_{i,i'}^T\}$. For a correlation dominant gene category C , the variance of U_W has a global maximum at $\delta_1 = \delta_2 = \dots = \delta_m = d$.*

Thus, the array permutation-based tests will tend to be conservative under realizations of (15) that depart from the special case in (13). This finding of conservativeness is also illustrated via the simulation study for categories that exhibit positive correlation on average without meeting the strict criterion of being correlation dominant (Figure 4D). We have also confirmed this result in a two-condition experiment simulated from a multivariate Gaussian model (not shown).

In the simulation above, the bootstrap methods maintained their approximately correct Type I error regardless of the Class 3 null hypothesis. However, in simulated expression data with a smaller sample size of $n = 20$ arrays, the anti-conservativeness of the quantile-based method becomes more pronounced at smaller target α . Since many microarray datasets can be of this size, the bootstrap Student’s t -interval is preferred.

6.3 Power under simulated alternatives

To assess the relative power of the bootstrap tests over array permutation, alternative hypotheses must be specified that relate to increased amounts of DE in a gene category. When the degree of DE can be expressed in terms of δ , an average increase within the category relative to its complement can be written as

$$H_A : \sum_{i=1}^K \beta_{C,k} \cdot d_k > \sum_{i=1}^K \beta_{\bar{C},k} \cdot d_k \quad (19)$$

For these alternatives the Wilcoxon rank sum, U_W , will be a global statistic well suited to identify increased amounts of differential expression in a robust manner.

In the randomized lung carcinoma dataset, realizations of (19) can be achieved by applying an additive or multiplicative constant to all gene-specific parameters within the category. More precisely, if $\{\delta_i^0 : i \in C\}$ are a category’s gene-specific parameters under the Class 3 null, we consider H_A to be of the form of either $\{\delta_i^A = c + \delta_i^0 : i \in C\}$

or $\{\delta_i^A = c \cdot \delta_i^0 : i \in C\}$. In this way, power curves can be displayed across a single axis by varying c . Figure 5 illustrates the effects when c is applied in an additive manner for $K = 2$ strata with DE and non-DE genes, and in a multiplicative manner for an example with $K = 5$ strata. The results demonstrate considerable improvements in power by the bootstrap methods over array permutation.

7 Analysis of a survival microarray dataset

The breast cancer survival datasets from Chang et al. (2005) is used to illustrate the power and utility of bootstrap resampling as compared to array permutation. A total of $n = 295$ breast cancer samples were analyzed on Agilent microarrays, and normalized gene expression estimates were obtained for a subset of $m = 11176$ genes that were annotated to at least one of 1348 GO terms (details on normalization, filtering, and formation of gene categories are omitted, but available from the authors). Survival times and censoring indicators were available for each array. Wald statistics from the univariate Cox proportional hazard model were used as local statistics to reflect the association between expression and patient outcome.

For the permutation- and bootstrap-based tests, the Wilcoxon rank-sum U_W was the global statistic with results obtained from 1000 resamples of the data. The p -values produced by the bootstrap quantile- and t -intervals were in good agreement across the set of categories (rank correlation > 0.999), reflecting that the distributions of resampled global statistics were nearly normally distributed. The permutation test also showed good agreement with the bootstrap (rank correlation of 0.977 with bootstrap results), but a distinct difference was observed in the number of categories achieving various levels of significance (Table 2). The improved power of the bootstrap

methods is apparent from the increased number of significant categories, with 48 declared significant via bootstrapping at $\alpha = 0.001$, but only 12 via permutation. Moreover, we have established that the increase in significant categories is far greater than could be produced by the slight anti-conservativeness of the bootstrap approach expected for this sample size. The minimal possible p -value of the permutation and bootstrap-quantile tests are limited by the 1000 resamples that were taken of the data. The bootstrap t -interval does not have this restriction, and 28 categories were observed to pass the conservative Bonferroni threshold for $\alpha = 0.05$. Because of the iterative procedure for estimates from the Cox-proportional hazard model, taking additional resamples of the dataset was computationally infeasible, and would be prohibitive when trying to control the FWER across such a large number of categories.

8 Discussion

We have used the terminology of local and global statistics as presented in SAFE (Barry et al. 2005) to describe existing methods for testing differential expression within a gene category. By classifying methods according to their assumed null hypotheses, we illustrate a number of shortcomings of these methods. We propose a novel bootstrap-based approach that uniquely allows both for genes within a category and its complement to be correlated, and that maintains proper error control under a more biologically sensible null hypothesis than has been implicitly used by other methods.

As a last but very important advantage to the bootstrap-based procedure, we note that by resampling with replacement, the bootstrap can incorporate covariate information in a sensible manner. In permutation testing, by inducing a null that

breaks the association between the response and expression, the covariate information can no longer be linked to both. Thus, a researcher is forced to choose the part of the data to remain linked to the covariate. By resampling the data jointly, the bootstrap allows the relationship between all three variable types to be maintained. The proper consideration of covariates is just one area of potential improvement, as gene category testing moves toward greater statistical maturity.

Appendix A: Proof to Theorem 1

The variance of U_W is decomposed using the Mann-Whitney form of the statistic.

$$\begin{aligned}
\text{Var}[U_W] &= \text{Var}\left[\sum_{i \in C} \text{Rank}(T_i)\right] \\
&= \text{Var}\left[\frac{m_C \cdot (m_C + 1)}{2} + \sum_{i \in C} \sum_{h \in \bar{C}} I\{T_i > T_h\}\right] \\
&= \sum_{i \in C} \sum_{i' \in C} \sum_{h \notin C} \sum_{h' \notin C} \text{Cov}[I\{T_i > T_h\}, I\{T_{i'} > T_{h'}\}] \tag{20}
\end{aligned}$$

where

$$\begin{aligned}
&\text{Cov}[I\{T_i > T_h\}, I\{T_{i'} > T_{h'}\}] \\
&= E[I\{T_i > T_h\} \cdot I\{T_{i'} > T_{h'}\}] - E[I\{T_i > T_h\}] \cdot E[I\{T_{i'} > T_{h'}\}] \\
&= P(\{T_h - T_i < 0\} \cap \{T_{h'} - T_{i'} < 0\}) - P(T_h - T_i < 0) \cdot P(T_{h'} - T_{i'} < 0)
\end{aligned}$$

and the pairs of differences follows a centered bivariate normal distribution

$$\begin{aligned}
&\begin{bmatrix} T_h - T_i \\ T_{h'} - T_{i'} \end{bmatrix} \sim N(\mathbf{0}, \Sigma) \quad \text{where} \tag{21} \\
\Sigma &= \begin{bmatrix} 2 - 2 \cdot \rho_{i,h}^T & \rho_{i,i'}^T + \rho_{h,h'}^T - \rho_{i,h'}^T - \rho_{i',h}^T \\ \rho_{i,i'}^T + \rho_{h,h'}^T - \rho_{i,h'}^T - \rho_{i',h}^T & 2 - 2 \cdot \rho_{i',h'}^T \end{bmatrix}.
\end{aligned}$$

Each term in (20) is evaluated as follows, where the pdf and cdf of a univariate and bivariate normal distribution are denoted by ϕ , Φ and ϕ_2 , Φ_2 respectively.

$$\begin{aligned}
&\text{Cov}[I\{T_i > T_h\}, I\{T_{i'} > T_{h'}\}] \\
&= \Phi_2\left(0, 0; \rho = \frac{\rho_{i,i'}^T + \rho_{h,h'}^T - \rho_{i',h}^T - \rho_{i,h'}^T}{\sqrt{(2 - 2\rho_{i,h}^T) \cdot (2 - 2\rho_{i',h'}^T)}}\right) - \Phi(0) \cdot \Phi(0) \tag{22}
\end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^0 \int_{-\infty}^0 \phi_2(x, y; \rho) dx dy - \frac{1}{4} \\
&= \int_{-\infty}^0 \int_{-\infty}^{\frac{-\rho^2 z_2}{\sqrt{1-\rho^2}}} \phi_2(z_1, z_2; \rho = 0) dz_1 dz_2 - \frac{1}{4} \tag{23}
\end{aligned}$$

$$= \int_0^\infty r \cdot \exp\left(-\frac{r^2}{2}\right) dr \cdot \int_\pi^{\frac{3\pi}{2} + \sin^{-1}(\rho)} \frac{1}{2\pi} d\theta - \frac{1}{4} \tag{24}$$

$$= \frac{1}{4} + \frac{\sin^{-1}(\rho)}{2\pi} - \frac{1}{4} = \frac{\sin^{-1}(\rho)}{2\pi} \tag{25}$$

where in (23) we have used the transformation $z_1 = (x - \rho y)/\sqrt{1 - \rho^2}$, $z_2 = y$ and then in (24) we used $z_1 = r \cos \theta$, $z_2 = r \sin \theta$.

Appendix B: Proof to Theorem 2

The following elementary lemma is useful in evaluating the expectation of U_W .

Lemma 1 *Let T_1 and T_2 be distributed as G_1 and G_2 and assume that $P(T_1 = T_2) = 0$. Define $\mu(G_1, G_2) \equiv E[I\{T_1 > T_2\}]$, then $\mu(G_1, G_2) = 1 - \mu(G_2, G_1)$ and $\mu(G_1, G_2) = 1/2$ when $G_1 = G_2$.*

The expectation of U_W is calculated by decomposing the $m_C \cdot m_{\bar{C}}$ pairwise comparison of T 's into the K^2 different terms involving $\mu(G_k, G_{k'})$.

$$\begin{aligned}
E[U_W] &= E\left[\sum_{i \in C} \text{Rank}(T_i)\right] = E\left[\frac{m_C \cdot (m_C + 1)}{2} + \sum_{i \in C} \sum_{h \notin C} I\{T_i > T_h\}\right] \\
&= \frac{m_C \cdot (m_C + 1)}{2} + \sum_{k=1}^K \sum_{k'=1}^K \sum_{\substack{i \in C \\ F_i = G_k}} \sum_{\substack{h \notin C \\ F_h = G_{k'}}} \mu(G_k, G_{k'}) \\
&= \frac{m_C \cdot (m_C + 1)}{2} + \sum_{k=1}^K \sum_{k'=1}^K m_C \cdot \beta_k \cdot m_{\bar{C}} \cdot \beta_{k'} \cdot \mu(G_k, G_{k'}) \\
&= \frac{m_C \cdot (m_C + 1)}{2} + m_C \cdot m_{\bar{C}} \left[\sum_{k=1}^K \frac{\beta_k^2}{2} + \sum_{k' < k} \beta_k \cdot \beta_{k'} \cdot [\mu(G_k, G_{k'}) + \mu(G_{k'}, G_k)] \right] \\
&= \frac{m_C \cdot (m_C + 1)}{2} + m_C \cdot m_{\bar{C}} \left[\sum_{k=1}^K \frac{\beta_k^2}{2} + \sum_{k' < k} \beta_k \cdot \beta_{k'} \right] \\
&= \frac{m_C \cdot (m_C + 1)}{2} + \frac{m_C \cdot m_{\bar{C}}}{2} \left[\sum_{k=1}^K \beta_k \right]^2 \\
&= \frac{m_C \cdot (m_C + 1)}{2} + \frac{m_C \cdot m_{\bar{C}}}{2} = \frac{m_C \cdot (m + 1)}{2}
\end{aligned}$$

Appendix C: Proof to Theorem 3

The following lemma regarding the bivariate normal distribution is useful for establishing an inequality for $\text{Var}[U_W]$.

Lemma 2 *For the bivariate normal distribution, the following is true for the function*

$$f(x, y) = \Phi_2(x, y; \rho) - \Phi(x) \cdot \Phi(y):$$

1. $f(0, 0)$ is a global maximum when $\rho > 0$
2. $f(0, 0)$ is a global minimum when $\rho < 0$
3. $f(x, y) = 0$ when $\rho = 0$

Proof: The first derivatives of $f(x, y)$ are

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) &= \frac{\partial}{\partial x} (\Phi_2(x, y; \rho) - \Phi(x) \cdot \Phi(y)) \\ &= \phi(x) \cdot \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) - \phi(x) \cdot \Phi(y) \end{aligned} \quad (26)$$

$$\propto \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) - \Phi(y) \quad (27)$$

and $\frac{\partial f}{\partial y}$ has an analogous form due to symmetry. Since Φ is a strictly increasing function, setting the derivatives equal to zero leads to the following equations

$$\begin{aligned} y - \rho x &= \sqrt{1 - \rho^2} \cdot y \\ x - \rho y &= \sqrt{1 - \rho^2} \cdot x \end{aligned} \quad (28)$$

for which $\{x = 0, y = 0\}$ is the only solution when $\rho \neq 0$. Since $(0, 0)$ is the only stationary point, a second derivative test can be used to determine whether it is a global minimum or maximum (Thomas and Finney 1992). The second derivatives

are solved to be

$$\begin{aligned}\frac{\partial f}{\partial x^2}(x, y) &= \phi'(x) \left[\Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) - \Phi(y) \right] + \phi(x) \cdot \phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \cdot \frac{-\rho}{\sqrt{1 - \rho^2}} \\ \frac{\partial f}{\partial y^2}(x, y) &= \phi'(y) \left[\Phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) - \Phi(x) \right] + \phi(y) \cdot \phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) \cdot \frac{-\rho}{\sqrt{1 - \rho^2}} \\ \frac{\partial f}{\partial x \partial y}(x, y) &= \phi(x) \cdot \left[\phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \cdot \frac{1}{\sqrt{1 - \rho^2}} - \phi(y) \right] = \frac{\partial f}{\partial y \partial x}(x, y)\end{aligned}$$

At the point $\{x = 0, y = 0\}$ the derivatives are equal to

$$\frac{\partial f}{\partial y^2}(0, 0) = \frac{\partial f}{\partial x^2}(0, 0) = \phi(0)^2 \cdot \frac{-\rho}{\sqrt{1 - \rho^2}} \quad (29)$$

$$\frac{\partial f}{\partial x \partial y}(0, 0) = \frac{\partial f}{\partial y \partial x}(0, 0) = \phi(0) \cdot \left[\phi(0) \cdot \frac{1}{\sqrt{1 - \rho^2}} - \phi(0) \right] \quad (30)$$

and the discriminant takes the form

$$\begin{aligned}D(0, 0) &= \frac{\partial f}{\partial x^2}(0, 0) \cdot \frac{\partial f}{\partial y^2}(0, 0) - \left(\frac{\partial f}{\partial x \partial y}(0, 0)\right)^2 \\ &= \left(\phi(0)^2 \cdot \frac{-\rho}{\sqrt{1 - \rho^2}}\right)^2 - \left(\phi(0) \cdot \left[\phi(0) \cdot \frac{1}{\sqrt{1 - \rho^2}} - \phi(0)\right]\right)^2 \\ &= \phi(0)^4 \left(\frac{\rho^2}{1 - \rho^2} - \frac{(1 - \sqrt{1 - \rho^2})^2}{1 - \rho^2}\right) \\ &= \phi(0)^4 \cdot 2 \cdot \frac{\sqrt{1 - \rho^2} - (1 - \rho^2)}{1 - \rho^2}\end{aligned} \quad (31)$$

Since $\sqrt{1 - \rho^2} > (1 - \rho^2)$ for all non-zero $\rho \in (-1, 1)$, the discriminant is strictly positive, proving that either a minimum or a maximum must exist. From the second derivatives in (29), one can show that $f(0, 0)$ is a minimum when $\rho < 0$ and a maximum when $\rho > 0$. Lastly $f(x, y)$ is exactly 0 when $\rho = 0$ by independence \square

The variance of U_W are decomposed into the covariances given in (20) as described in Theorem 1, but unlike (21), the paired differences in local statistics follow non-central bivariate normal distributions under (15) with marginal means $\delta_h - \delta_i$ and

$\delta_{h'} - \delta_{i'}$. From (22) each covariance term can be written as

$$\Phi_2\left(\delta_h - \delta_i, \delta_{h'} - \delta_{i'}; \rho\right) - \Phi(\delta_h - \delta_i) \cdot \Phi(\delta_{h'} - \delta_{i'}) \quad (32)$$

where ρ is defined as in (22). We consider in turn several cases.

When $i = i'$ and $h = h'$, ρ is proportional to $2 - 2 \cdot \rho_{i,h}^T$, which is positive quantity except when the genes are perfectly correlated which is ruled out by the definition of a correlation dominant category. From Lemma 2, (32) is maximized when $\delta_i = \delta_h$. Since this is true for all $\{i, h\}$ pairs of category and complement genes, a global maximum of the summed covariances will occur when all local statistics have the same mean.

When $i = i'$ and $h \neq h'$, ρ is proportional to $1 + \rho_{i,i'}^T - \rho_{i',h}^T - \rho_{i,h'}^T$ and will be greater than 0 for a correlation dominant category such that a maximum occurs when $\delta_i = \delta_h = \delta_{h'}$. An analogous argument holds for when $i \neq i'$ and $h = h'$.

For $i \neq i'$ and $h \neq h'$, either ρ will be positive if $(\rho_{i,i'}^T + \rho_{h,h'}^T) > (\rho_{i',h}^T + \rho_{i,h'}^T)$ so that (32) is maximized when $\delta_h = \delta_i$ and $\delta_{h'} = \delta_{i'}$, or ρ will be exactly 0 if $(\rho_{i,i'}^T + \rho_{h,h'}^T) = (\rho_{i',h}^T + \rho_{i,h'}^T)$ and (32) will be constant. This inequality of summed correlations is again guaranteed for correlation dominant categories.

This proves a global maximum for $\text{Var}[U_W]$ is achieved at $\delta_1 = \delta_2 = \dots = \delta_m = d$ since only in this case will every covariance term in (20) be either maximized, or a constant.

References

- S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.
- V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, 2001.
- M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.
- M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and Sherlock G. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25–29, 2000.
- K. I. Virtaneva, F. A. Wright, S. M. Tanner, B. Yuan, W. J. Lemon, M. A. Caligiuri, C. D. Bloomfield, A. de la Chapelle, and R. Krahe. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci U S A*, 98(3):1124–1129, 2001.
- William T. Barry, Andrew B. Nobel, and Fred A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.
- Seon-Young Kim and David J Volsky. Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005.
- A. Boorsma, B. C. Foat, D. Vis, F. Klis, and H. J. Bussemaker. T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Research*, 33:W592–W595 Suppl. S JUL 1 2005, 2005.
- V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273, 2003.
- Y. Ben-shaul, H. Bergman, and H. Soreq. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21:1129–1137, 2005.
- D. B. Allison, X. Q. Cui, G. P. Page, and et al. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65, 2006.
- S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–04, 2003.

- T. Beißbarth and T. P. Speed. Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
- K. Pearson. On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8:250–254, 1911.
- D. Damian and M. Gorfine. Statistical concerns about the GSEA procedure. *Nature Genetics*, 36:663–663, 2004.
- S. Zhong, K. F. Storch, O. Lipan, M. C. Kao, C. J. Weitz, and W. H. Wong. GoSurfer : A graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl Bioinformatics*, 3(4):261–264, 2004.
- P. Pavlidis, J. Qin, V. Arango, J. J. Mann, and E. Sibille. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, 29:1213–1222, 2004.
- A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, 98(24):13790–13795, 2001.
- H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14:1085–1094, 2004.
- T. Galitski, A. J. Saldanha, C. A. Styles, E. S. Lander, and G. R. Fink. Ploidy regulation of gene expression. *Science*, 285(5425):251–254, 1999.
- George Casella and Roger L. Berger. *Statistical inference*. Duxbury, Australia, second edition, 2002.
- A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102:15545–15550, 2005.
- Brad Efron and Rob Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, in press.
- B Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, New York, New York, second edition, 1998.
- B Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.
- B. Efron. Better bootstrap confidence-intervals. *Journal Of The American Statistical Association*, 82:171–185, 1987.

- H. Y. Chang, D. S. A. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, H. Y. Dai, Y. D. D. He, L. J. Van't veer, H. Bartelink, M. Van de rijm, P. O. Brown, and M. J. Van de vijver. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102:3738–3743, 2005.
- G. B. Jr. Thomas and R. L. Finney. *Maxima, Minima, and Saddle Points*. Calculus and Analytic Geometry. Addison-Wesley, Reading, MA, eighth edition, 1992.

		Significant DE			
		Yes	No		
Category, C	$\sum_{i \in C} I\{T_i \in \Gamma\}$			m_C	
Complement, \bar{C}				$m_{\bar{C}}$	
		r	$m - r$	m	

Figure 1: Results from a gene-specific analysis provided in a 2×2 table for a category versus its complement. The size of the two gene sets, given by m_C and $m_{\bar{C}}$ respectively, are assumed to be fixed quantities. The complete table is then determined by knowing the number of rejections in the category, and either the number of rejections in the complement or fixing the total number of rejections, $R = r$.

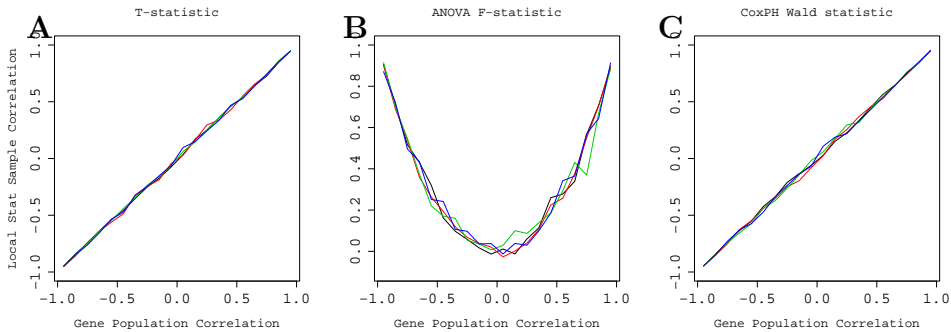


Figure 2: Correlations in expression and local statistic from Monte Carlo simulations of normally distributed expression for two genes in several choices of experimental design. **(A)** Student's t for a two-sample comparison; **(B)** F statistic for an ANOVA with 4 groups; **(C)** Cox-proportional hazard model for relating expression to exponentially distributed survival and censoring times. In each design, the variance of expression in the second gene was 1, 2, 5 and 10 fold greater. Data was simulated for $n = 40$ arrays with equal sample sizes per group.

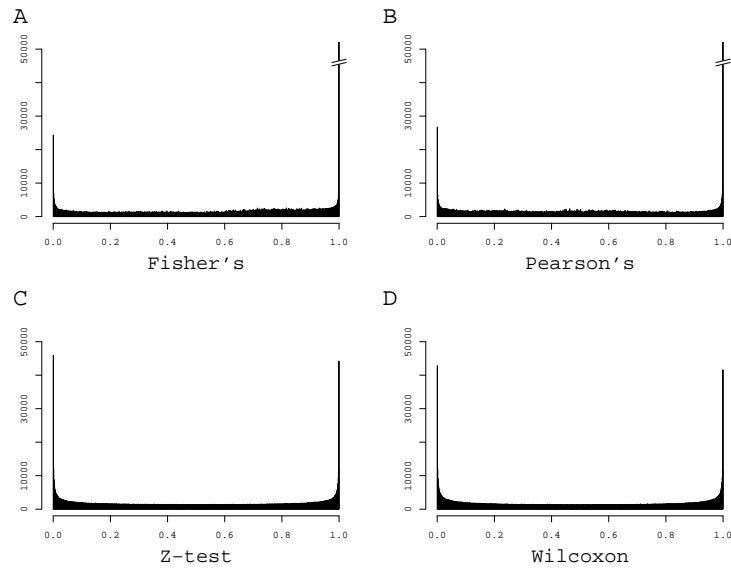


Figure 3: Histograms of p -values (1823 categories over 1000 simulations) for the gene-list enrichment tests: **(A)** Fisher's Exact Test and **(B)** and Pearson's difference in proportions; and for the continuous tests of: **(C)** Average difference and **(D)** Wilcoxon rank sum. The large number of small and large p -values demonstrate the over dispersion that results from incorrect variance estimates.

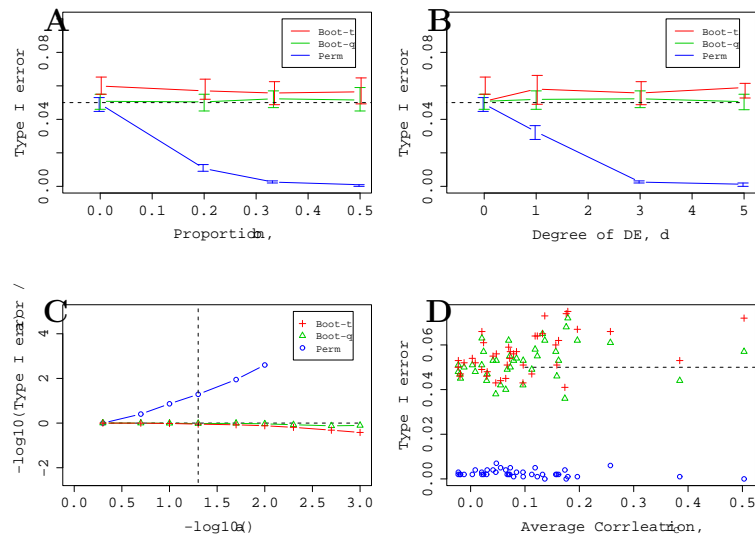


Figure 4: Performance of bootstrap- and permutation-based SAFE tests under different null hypotheses. The average Type I error of a category is shown for (A) four different proportions of DE and (B) for four different levels of DE. (C) The Type I error at different α levels is shown for $d = 3$ and $\beta = 1/3$, and (D) is plotted for each category against the average pairwise correlation.

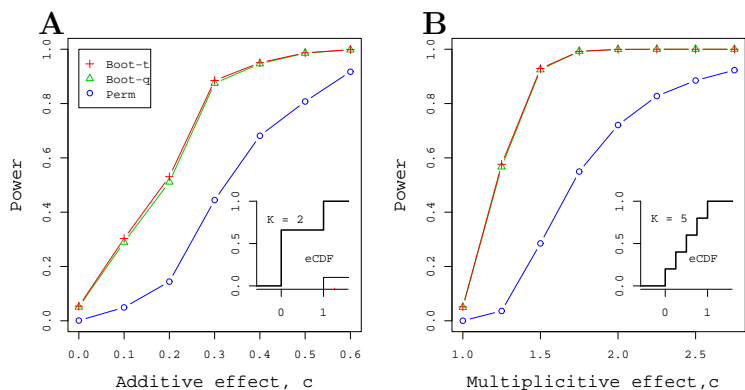


Figure 5: Average power of permutation and bootstrap based gene category tests as one departs from Class 3 nulls. Results based on randomized microarray data and real GO categories, and applying (A) an additive constant to $K = 2$ classes of genes with $1/3$ differentially expressed at $d = 1$ (as shown by the CDF in the inset graphs), and (B) a multiplicative constant to $K = 5$ classes of genes with $\{d_k\}$ equally spaced between 0 and 1. Both scenarios exhibit more power to detect the alternative using the bootstrap tests.

Table 1: The ratio of realized Type I error rates to target α levels.

	Fisher, U_F	Pearson, U_P	Ave Diff, U_D	Wilcoxon, U_W
Class 1 tests				
$\alpha = 0.1$	1.19	1.32	1.82	1.86
$\alpha = 0.01$	3.40	3.48	5.92	5.83
$\alpha = 0.001$	13.4	14.7	25.2	23.5
$\alpha = 1e - 4$	65.6	72.5	130	116
$\alpha = 1e - 5$	367	431	769	677
$\alpha = 1e - 6$	2213	2922	4974	4245
Class 2 tests				
$\alpha = 0.1$	0.39	1.01	1.01	1.01
$\alpha = 0.01$	0.21	1.01	1.01	1.01
$\alpha = 0.001$	0.14	1.05	1.03	1.01
$\alpha = 1e - 4$	-NA-	-NA-	-NA-	-NA-

Table 2: Number of significant GO categories for target α levels.

	Perm	Boot-q	Boot-t
$\alpha = 0.1$	195	222	220
$\alpha = 0.05$	129	157	160
$\alpha = 0.01$	56	72	85
$\alpha = 0.005$	36	63	73
$\alpha = 0.001$	12	40	48
$\alpha = 3.7e - 5^*$	-NA-	-NA-	28

* Bonferroni cutoff