Biostatistics (2013), **14**, 3, *pp.* 573–585
doi:10.1093/biostatistics/kxt004
Advance Access publication on February 20, 2013

Empirical pathway analysis, without permutation

YI-HUI ZHOU*
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
yihuiz@live.unc.edu
WILLIAM T. BARRY
Dana-Farber Cancer Institute, Boston, MA 02215, USA
FRED A. WRIGHT
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

Summary

Resampling-based expression pathway analysis techniques have been shown to preserve type I error rates, in contrast to simple gene-list approaches that implicitly assume the independence of genes in ranked lists. However, resampling is intensive in computation time and memory requirements. We describe accurate analytic approximations to permutations of score statistics, including novel approaches for Pearson's correlation, and summed score statistics, that have good performance for even relatively small sample sizes. Our approach preserves the essence of permutation pathway analysis, but with greatly reduced computation. Extensions for inclusion of covariates and censored data are described, and we test the performance of our procedures using simulations based on real datasets. These approaches have been implemented in the new *R* package *safeExpress*.

Keywords: Gene sets; Multiple hypothesis testing; Permutation approximation.

1. Introduction

A basic approach to gene expression analysis involves the detection of genes significantly associated with clinical outcome, treatment, or experimental design variables. This kind of analysis focuses on individual genes. However, researchers are often interested in the association of the clinical/treatment variable with sets of genes of related biological function, either to increase power or to provide a more parsimonious, set-based view of the results. We use the term *pathway* to refer generically to a gene set under study, whether or not the set is a true metabolic or signaling pathway. Numerous methods for pathway analysis have been proposed (Dinu *and others*, 2009), and can be divided into approaches that implicitly assume uncorrelated expression data vs. those that acknowledge correlation (Barry *and others*, 2008). As described in Gatti *and others* (2010), approaches that acknowledge correlation via permutation have vastly superior type I error control properties compared with methods that assume no correlation, and include Gene Set Enrichment Analysis (GSEA) (Subramanian, 2005) and Significance Analysis of Function and Expression (SAFE) (Barry *and others*, 2005). The *globaltest* of Goeman *and others* (2004) offers a non-resampling approach, in which the correlation structures are parametrically modeled.

*To whom correspondence should be addressed.

The SAFE procedure computes *local* statistics, which measure the association between individual genes and the clinical/treatment variable, and *global* statistics, which are aggregations of local statistics for an entire pathway. Local statistics can be directional, i.e. reflecting positive or negative association with the clinical/treatment variable, or non-directional, in which only the magnitude of association is of interest. Global statistics can be self-contained, meaning that local statistics are aggregated within the pathway only, or competitive, comparing the genes in the pathway to those in the complement (following the terminology of Goeman and Buhlmann, 2007). A wide variety of procedures, including GSEA, are encompassed in the framework, simply by varying the statistics (Barry *and others*, 2008). In this paper, we focus on score-based local statistics and global statistics based on sums and differences. No further familiarity with the literature on pathway analysis is assumed.

Gene set analysis using permutation of samples dates to Virtaneva *and others* (2001), in which the clinical/treatment variable is permuted relative to the expression data. The appeal of permutation is that it conditions on gene expression without requiring an explicit understanding of the underlying structure. As described in Gatti *and others* (2010), proper handling of the correlation of gene expression levels among genes in a set is essential for error control. Once proper per-pathway p-values are obtained, standard techniques for controlling the false discovery rate or family-wise error rate across multiple pathways may be used.

A downside to permutation is that it is computationally intensive, keeping in mind that all genes must be examined for each permutation (for competitive global statistics). When testing numerous gene sets (which may reach several thousand), it may be difficult to achieve multiple test-corrected significance, unless the number of permutations is very large. Using standard desktop computing, performing 1000 permutations of the entire dataset is typical (Knijnenburg *and others* (2009)). Effective false-positive control across large numbers of pathways may not be possible, as the standard empirical p-values are constrained to a minimum 1/(number of permutations).

As an alternative viewpoint, we note that correlation among local test statistics is induced by the correlation of expression levels among genes (Barry *and others*, 2008). It is thus worth exploring whether the permutation null distributions of global statistics can be parametrically approximated, using empirical estimates of correlation structure. We emphasize that actual permutation, although unbiased, can be subject to considerable sampling variability unless the number of permutations is large. Here, we provide new analytic approximations, with comparisons with random permutation. For a proposed non-directional competitive global statistic, we are not aware that competing parametric approximations have been proposed.

2. METHODS

2.1 *Notation*

We suppose that each sample $j = 1, 2, \ldots, n$ is associated with a clinical/treatment value x_j, with the vector denoted by $\mathbf{x}$, and can be discrete or continuous. We later describe approximate procedures to handle censored time-to-event data. We use g_{ij} to denote the expression level for the ith gene ($i = 1, 2, \ldots, m$) and jth sample. Let $\mathbf{Y}$ be the $m \times n$ normalized gene expression matrix, with $y_{ij} = (g_{ij} - \overline{g_{i.}})/\sqrt{\sum_{j=1}^{n}(g_{ij} - \overline{g_{i.}})^2/n}$, where $\overline{g_{i.}} = \sum_j g_{ij}/n$. This normalization produces $\sum_{j=1}^{n} y_{ij} = 0$ and $\sum_{j=1}^{n} y_{ij}^2 = 1$. An important local statistic to represent the relationship between the ith gene and the clinical/treatment variable is the score statistic, which for linear regression and a variety of generalized linear models can be expressed as (Schaid and Sommer, 1994)

$$S_i = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}_{i.}}{\sqrt{\Sigma_{j=1}^{n}(x_j - \bar{x})^2/n}} = \frac{\sum_j x_j y_{ij}}{s_x\sqrt{(n-1)/n}}, \tag{2.1}$$

where s_x is the sample standard deviation of the $\mathbf{x}$ values. Our formulation follows that of Lee *and others* (2011) (except for notational differences), who obtained results for score statistics that we use below, although in a different context. Score statistics have comparable asymptotic power properties as Wald and likelihood ratio statistics for local departures from the null, although in small-sample settings other statistics may have slight advantages (Harris and Peers, 1980). One source of confusion in assessing power arises from improper control of type I error rates, and we are motivated to consider exact testing (corresponding to exhaustive permutation) as a "gold standard". We note that, for each S_i, the roles of $\mathbf{y}_{i\cdot}$ vs. $\mathbf{x}$ are interchangeable, and the same score statistic applies to a number of models in which $\mathbf{x}$ is treated as the response variable and gene expression $\mathbf{Y}$ as the predictor. Examples include logistic or Poisson regression for binary or integer $\mathbf{x}$.

2.1.1 *Survival analysis.* For right-censored clinical data, we use the martingale residuals

$$x_j = \delta_j - \hat{\Lambda}_0(t_j),$$

(Therneau and Grambsch, 1990), where δ_j is the death (i.e. non-censored) indicator for the jth individual at the observed time t_j and $\hat{\Lambda}_0$ is the estimated cumulative hazard. Martingale residuals have also been used in a version of $globaltest$ (Goeman *and others*, 2005), and provide considerable computational convenience compared with proportional hazards regression. As $\mathbf{x}$ and $\mathbf{Y}$ are treated symmetrically in the score statistic, it does not matter that the censored data are more naturally considered a response value than a predictor.

2.2 *Self-contained testing*

Under the complete null hypothesis that no gene shows differential expression, it is reasonable to focus on each pathway while ignoring the remaining genes. We use $\{path\}$ to denote a pathway with m_{path} genes. We (Barry *and others*, 2008) have criticized self-contained testing for failing to account for gene effects not specific to a pathway. Nonetheless, self-contained testing may sometimes be reasonable. These include situations where (i) few genes are differentially expressed; (ii) no gene is significant when accounting for genome-wide multiple comparisons; or (iii) a candidate pathway is tested, where any evidence of association is of interest.

2.2.1 *The directional statistic U.* A simple aggregation of directional local statistics S_i is $U = \sum_{i\in\text{path}} S_i$, which is sensitive to associations between expression and the clinical/treatment variable that are in the same direction. Nonetheless, testing for this statistic will generally be two-sided. Motivated by the central limit theorem, a normal approximation to U might seem appealing, as summation is performed both over n (within S_i) and m_{path}. However, the distribution of U is fundamentally limited by n, shown by rewriting

$$U = \frac{\sum_{i\in\text{path}}\sum_j x_j y_{ij}}{s_x\sqrt{(n-1)/n}} \propto \sum_j x_j \left(\sum_{i\in\text{path}} y_{ij}\right) = \sum_j x_j y'_j,$$

for $y'_j = \sum_{i\in\text{path}} y_{ij}$ (with the vector of these values denoted by $\mathbf{y}'$). In this formulation, it is simple to show (Appendix A1 of supplementary material available at *Biostatistics* online) that under permutation U is one-to-one with the Pearson correlation r_U between $\mathbf{y}'$ and permutations of $\mathbf{x}$. This immediately suggests the density approximation for r_U^2 as Beta$\left(\frac{1}{2}, \frac{1}{2}(n-2)\right)$, the exact unconditional null density if either $\mathbf{x}$ or $\mathbf{y}'$ is drawn from a normal density. Under permutation, the expectation of r_U^2 is the same as that predicted by the standard beta approximation. However, the

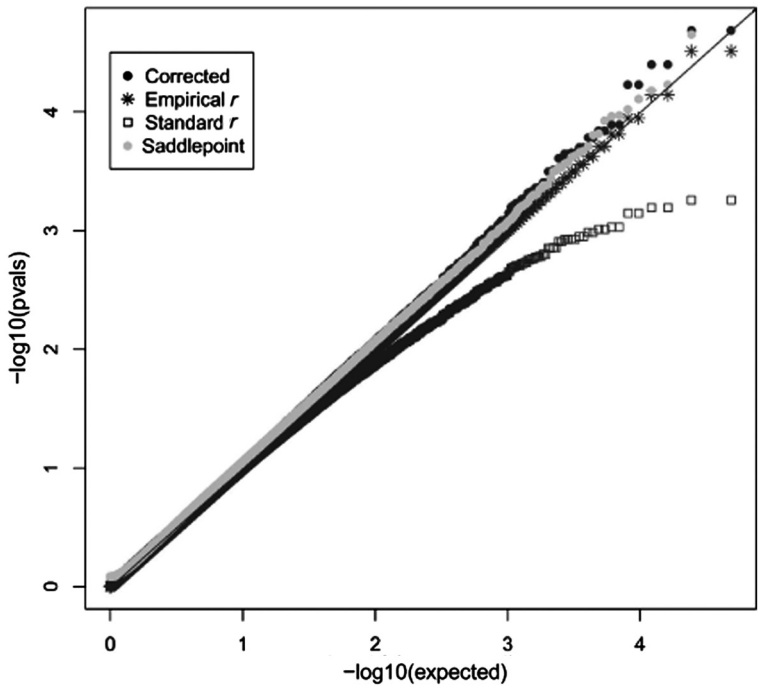

Fig. 1. Qqplots of p-values for r_U^2, computed over exhaustive permutations on the salivary gland data, for the 46 probesets in the KEGG:00271 pathway ($n_0 = n_1 = 9$).

permutation variance may differ from the standard approximation. Here, we improve the standard approximation by choosing beta parameters to match the first two exact permutation moments of r_U^2 (Appendix A1 of supplementary material available at *Biostatistics* online).

2.2.2 *An example.* As a simple example where the standard beta approximation can fail, we consider a salivary gland expression dataset (ArrayExpress E-GEOD-7451, Affymetrix U133 plus 2.0). Of 20 original arrays, two appeared to be accidental duplicates, and we restrict attention to the remaining 18, of which $n_1 = 9$ with Sojgren's syndrome were compared with $n_0 = 9$ controls, and with the patient age reported. For the 46 probe sets (essentially genes) in pathway 00271 from the Kyoto Encyclopedia of Genes and Genomes (KEGG:00271), Methionine metabolism, the empirical r_U^2 p-values were computed for the $\binom{18}{9} = 48\,620$ unique permutations, and are guaranteed to be rank-uniform. Figure 1 shows a qqplot for the various p-values approaches, along with p-values provided by a saddlepoint approximation for the dichotomous setting (Appendix A2 of supplementary material available at *Biostatistics* online). The corrected Beta has similar performance as the saddlepoint, but is much simpler to calculate, and its underlying approximation is easily incorporated into other global statistics described below.

2.2.3 *The non-directional statistic V.* The squared score (local) statistics are non-directional, and a simple non-directional global statistic to detect departures from the self-contained null is a summation of the squared score statistics in the pathway,

$$V = \sum_{i \in \text{path}} S_i^2. \tag{2.2}$$

This statistic is widely used in association testing of single nucleotide polymorphisms (SNPs) and genes (Liu, 2010). Moreover, for the expression context it is equivalent to the *globaltest* (Goeman *and others*, 2004) when the expression data have been scaled for each gene (Pan, 2009). Goeman *and others* (2004) argued that V is optimal when aggregating small effect sizes. A standard approximation uses quadratic form results and moment-matching based on non-central (Liu *and others*, 2009), or scaled (Duchesne and Micheaux, 2010) χ^2. "Naive" moment estimates here are $E_\Pi(V) \approx m_{\text{path}}$ (which would follow from a χ^2_1 approximation for each S_i^2), and $\text{var}_\Pi(V) \approx 2\,\text{trace}(R^{\text{T}}R)$, where $R = \mathbf{Y}_{\text{path}}^{\text{T}}\mathbf{Y}_{\text{path}}$ is the correlation matrix for the genes in {path}. However, for small to moderate sample sizes, approaches described in Goeman *and others* (2004) can be used to correct the moments. We highlight an alternate formulation, described, for instance, in Lee *and others* (2011):

$$V = n\sum_{j=1}^{n-1}\lambda_j r_j^2, \tag{2.3}$$

where λ_j is the jth eigenvalue of $\mathbf{Y}_{\text{path}}^{\text{T}}\mathbf{Y}_{\text{path}}$, and r_j^2 is the squared Pearson correlation between the jth principal component of $\mathbf{Y}_{\text{path}}$ and $\mathbf{x}$ (for $j = 1, \ldots, n-1$). In other words, $r_j = \text{corr}(\mathbf{p}_{\cdot j}, \mathbf{x})$, where $\mathbf{p}_{\cdot j}$ is the jth column of $\mathbf{P}$ and $\mathbf{P}^{\text{T}}$ is the right singular matrix in the singular value decomposition of $\mathbf{Y}_{\text{path}}$.

Equation (2.3) shows how the rank of $\mathbf{Y}_{\text{path}}$ affects V. If $m_{\text{path}} < n-1$, some of the λ_j will be zero, and, even for large m_{path}, the number of contributing terms cannot exceed $n-1$. Note that the orthogonality of the first $n-1$ principal components (PCs) (for pathways with $m_{\text{path}} \geqslant n-1$) implies that $\sum_j r_j^2 = 1$, while, for constant V, (2.3) is the equation of an ellipse. An elementary comparison of these two expressions (a circle vs. an ellipse where each r_j is a coordinate) implies that V cannot exceed $n\lambda_1$. This is among the reasons that χ^2-based approximations to V can fail in the extreme tail, even if the moments are specified correctly.

We use (2.3) to motivate an alternative analytic approximation to the permutation distribution of V. The standard r^2 approximation implies that each $r_j^2 \sim \text{Beta}\left(\frac{1}{2}, (n-2)/2\right)$, noting that any of the PCs can serve the role played by $\mathbf{y}'$ in the earlier subsection. Although the PCs are orthogonal, the $\{r_j^2\}$ are actually somewhat negatively correlated over permutation, as implied by results in Appendix B2 of supplementary material available at *Biostatistics* online. Unconditional multivariate normality assumptions for $\mathbf{x}$ and $\mathbf{P}$ imply that the $\{r_j^2\}$ follow a correlated joint beta density, related to the multiple correlation coefficient sampling distribution (Fisher, 1938). We modify this density for our permutation setting, where normality is not assured, by replacing the standard beta parameters with the exact moment-corrected versions, as derived earlier for r_U^2. Examples below show that the approach works well, even for dichotomous $\mathbf{x}$ and modest sample sizes. The approximating joint density is derived in Appendix B1 of supplementary material available at *Biostatistics* online, but may be simply described as a recursion of successive r_j^2. Recalling that $r_j = \text{cor}(\mathbf{p}_{\cdot j}, \mathbf{y})$, we have the standard approximation $r_1^2 \sim \text{Beta}\left(\frac{1}{2}, (n-2)/2\right)$, and show that, for any subset $\Omega \subset \{1, \ldots, n-1\}$ that does not contain k, $r_k^2/(\{1 - \sum_{j\in\Omega} r_j^2\}) \sim \text{Beta}\left(\frac{1}{2}, (n - |\Omega| - 2)/2\right)$. Here $|\Omega|$ denotes the number of elements in Ω. Therefore, if r_1^2 is generated, the remaining values can be drawn conditionally as $r_2^2 = B_1 \times (1 - r_1^2)$, $r_3^2 = B_2 \times (1 - r_1^2 - r_2^2)$, etc., where each B_j is an independent draw from $\text{Beta}\left(\frac{1}{2}, (n-1-j)/2\right)$. Thus, V can be approximated as a sum of weighted correlated beta variates. The approximation reflects correlation structure that may be non-trivial for moderate sample size, with tails that are short enough to be realistic. The recursion applies to any ordering of the PCs, but ordering by eigenvalues is helpful for numeric approximation.

We estimate p-values by numeric integration over a 2D grid of $\{r_1^2, r_2^2\}$ for the initial term $\lambda_1 r_1^2 + \lambda_2 r_2^2$ (100×100, or 1000×1000 for more accuracy). Then a shifted gamma approximation is used for the remaining terms in (2.3) (Appendix B2 of supplementary material available at *Biostatistics* online). The shifted gamma distribution is an ordinary gamma with an additional location parameter, specified by

moment-matching from the eigenvalues. Our procedure enables testing of thousands of categories on a standard PC.

Equation (2.3) is similar to the asymptotic representation of Goeman and van Houwelingen (2011) for the globaltest statistic as a weighted sum of independent χ^2 random variables, providing important large-sample motivation for our approximations below in the generalized linear model setting. In addition, the authors provide an alternate representation under the linear model for finite samples. We emphasize that our representation in (2.3) is exact over permutations, thus transparently highlighting the importance of the PCs.

2.3 *Competitive testing*

Competitive global test statistics contrast the local statistics within each pathway vs. those of the complementary gene set (Goeman and Buhlmann, 2007). For datasets in which many genes are associated with $\mathbf{x}$, this approach enables a focus on truly pathway-related findings, rather than non-specific results that apply to all genes. Note that permutation induces a special case of the null hypothesis in which all genes are null, but competitive tests remain interpretable, reflecting the correlation structure in each pathway vs. its complement. Barry *and others* (2008) discuss these issues in detail, showing that permutation may be slightly conservative when the competitive null holds, i.e. genes may exhibit differential expression, but the pathway is not enriched for such genes.

For aggregations of directional local statistics, a straightforward competitive global statistic is $U_{\text{path}}/m_{\text{path}} - U_{\text{comp}}/m_{\text{comp}}$, where {comp} designates the complementary set of genes not in {path}. However, for array studies we must consider the role of data normalization, and we note that $U_{\text{all}} = U_{\text{path}} + U_{\text{comp}}$. If the data have been normalized so that each array has the same mean expression (which is true for simple centering procedures or for quantile normalization), then U_{all} will be constant, and U_{path} and U_{comp} will have correlation -1 over permutations. Thus, the construction of a competitive global statistic would be redundant, and U_{path} is essentially *already* a competitive statistic. Although this statement is not strictly true for other normalization procedures, it seems clear that normalization is likely to have a large impact on inference, and we argue that the original U_{path} should not be further modified in an attempt to make it "competitive".

2.3.1 *The competitive statistic D.*

A natural competitive global statistic based on sums of squared score statistics is $D = V_{\text{path}}/m_{\text{path}} - V_{\text{comp}}/m_{\text{comp}}$. As m_{comp} is typically much larger than m_{path}, one might expect that $V_{\text{comp}}/m_{\text{comp}}$ is nearly constant. In fact, for many datasets, the gene–gene correlation structures are sufficiently strong that variation in $V_{\text{comp}}/m_{\text{comp}}$ is non-negligible. Moreover, V_{path} and V_{comp} are correlated. The primary barrier to approximating the distribution of D is that the form of (2.2) cannot be negative, involving sums of squared terms, while its equivalent (2.3) is critical to the beta-mixture approach. As we show in detail in Appendix C of supplementary material available at *Biostatistics* online, this obstacle is neatly overcome by adding weights w_i to the score statistics $\{S_i^2\}$, where the weights for $i \in \{\text{comp}\}$ are imaginary, resulting in negative squared terms. Formally, we create a new weighted matrix $\mathbf{A} = \mathbf{WY}$, where $\mathbf{W}$ is the diagonal matrix with terms $\{w_i\}$. Here, a new equivalent relation holds,

$$D = \sum_{i=1}^{m} w_i^2 S_i^2 = n \sum_{j=1}^{n} \gamma_j c_j^2, \tag{2.4}$$

where $\{\gamma_j\}$ are the eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, and c_j^2 are the observed squared correlations between the corresponding eigenvectors and $\mathbf{x}$. We see that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is a real matrix, and so no complex algebra is required for the decomposition. Moreover, computation is greatly simplified by computing $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ once, and computing

$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}_{\mathrm{path}}^{\mathrm{T}}\mathbf{Y}_{\mathrm{path}}$ for each pathway. Equation (2.4) shows that D may be approximated by the beta mixture, but with weights γ_j that are both positive and negative (and in fact sum to zero).

An alternative competitive statistic based on ratios, rather than differences, is also described in Section 3.

2.4 *Inclusion of covariates*

Suppose $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p$ are a set of n-vector covariates, any of which may be correlated with $\mathbf{Y}$ or $\mathbf{x}$, or perhaps both. In principle, score statistics in the presence of covariates involve straightforward maximization over a restricted null space, which could be applied for each gene. However, we still need to handle gene–gene correlation, for which permutation is attractive. The proper handling of covariates is a challenge in the permutation setting (Buzkova *and others*, 2011). We employ one of the approaches described in Kennedy and Cade (1996), computing new local statistics $S_{i,z} = \sum_{j=1}^{n} x_{j,z} y_{ij,z} / \sqrt{\sum_j ((x_{j,z} - \bar{x}_z)^2/n)}$, where $\mathbf{Y}_{i.,z}$ and $\mathbf{x}_z$ have been adjusted by linear regression for the $n \times p$ covariate matrix $\mathbf{Z}$. Thus, $S_{i,z}$ is proportional to the partial correlation between $\mathbf{y}_{i.}$ and $\mathbf{x}$ after adjusting for $\mathbf{Z}$. Then we compute U_z, V_z, and D_z using the methods described earlier. However, we do employ a correction for the effective loss of p degrees of freedom (Appendix D of supplementary material available at *Biostatistics* online).

Collectively, we refer to all of the proposed methods above as the *safeExpress* procedure, with an accompanying R package.

3. Results

3.1 *Power*

A number of studies have compared the power of various local/global tests. The connection of V to the widely used globaltest suggests that it should have high power for a range of alternatives. Moreover, U and D are quite simple in their justification and connection to V and to the underlying score statistics. Liu *and others* (2007) and Dinu *and others* (2007) investigated a number of approaches and found that the globaltest was highly competitive, and several authors have discussed weaknesses in the GSEA Kolmogorov–Smirnov statistic (Barry *and others*, 2008; Dinu *and others*, 2007). The study of Fridley *and others* (2010) examined 10 different self-contained global statistics for a continuous $\mathbf{x}$. The authors found that the Fisher combined p-value (FCP, for which permutation was necessary to evaluate significance) was the most powerful practical statistic, essentially equal in power to more complex statistics.

For each statistic and scenario, Fridley *and others* (2010) performed 1000 simulations, and 1000 permutations for each to obtain empirical p-values. In addition to the large number of statistics considered, these authors' efforts are notable for the range of scenarios, varying the size of the pathway (10, 50, 100 genes), sample size ($n = 20$, 50, 100, 500), and average gene–gene correlation ($\rho = 0$, 0.1, 0.3), with each test performed at intended $\alpha = 0.05$. Effect sizes were determined by simulating the continuous phenotype as $x_j = \sum_i \beta_i y_{ij} + \epsilon$ for a coefficient vector $\boldsymbol{\beta}$ consisting of zeros and positive values, and normally distributed ϵ.

In all, 108 null and 2160 alternative hypothesis scenarios were simulated. Following the authors' simulation protocol, we ran *safeExpress* for 1000 simulations of each scenario. Figure S1 of supplementary material available at *Biostatistics* online shows the type I error rates for the null scenarios. All of our statistics and approximations control type I error at the expected rate. For the alternative scenarios, the results confirm those reported by Fridley *and others* (2010), and are highly supportive of our approaches (Figure S2 of supplementary material available at *Biostatistics* online). Averaging across scenarios for each sample size, the power of U was somewhat less (4% points) than that of the FCP (Figure S2(A) of

Table 1. *Performance of the p-value approximations for KEGG:00940* (18 *genes*), *salivary gland dataset* ($n = 18$)

Threshold	Proposed approximation					
	Continuous **x**			Dichotomous **x**		
α	U	V	D	U	V	D
10^{-1}	1.00	0.93	0.96	1.00	0.92	0.94
10^{-2}	1.00	0.92	0.94	1.01	0.97	0.88
10^{-3}	1.01	0.90	0.93	0.95	0.69	0.84
10^{-4}	1.02	0.89	0.90	1.23	1.22	1.22
10^{-5}	1.04	0.97	1.00	—	—	—
10^{-6}	1.11	0.59	0.71	—	—	—

Entries in the table are ratio$_\alpha$ = (true type I error rate)/α. For dichotomous **x** ($n_0 = n_1 = 9$), p-values are limited by the number of possible permutations.

supplementary material available at *Biostatistics* online). However, for some of the scenarios, U was more powerful, primarily with small samples ($n = 20$), low correlation ($\rho = 0$), and large cumulative effect sizes $\sum_i \beta_i$ (Figure S2(B) of supplementary material available at *Biostatistics* online). The power of the FCP was nearly identical to that of V and D (Figures S2(C) and (D) of supplementary material available at *Biostatistics* online). For D, we simulated an additional 10 000 genes in the complementary gene set with the same common ρ specified for the gene set. The complete set of null and power results are presented as supplementary files. While these simulations are extensive, power analyses presented elsewhere have already consistently supported simple aggregations of linear statistics. The overall conclusion is clear. Our aggregations of score statistics have competitive power with the most powerful statistics proposed by others, but for which significance is assessed by permutation.

Finally, as a possible alternative to D, we assessed the power of a competitive statistic $D_{\text{ratio}} = (V_{\text{path}}/m_{\text{path}})/(V_{\text{comp}}/m_{\text{comp}})$, i.e. using ratios instead of differences. Here the significance of D_{ratio} was determined by direct permutation, as we are not aware of an effective approximation. Within each of FCP power values near 0.1, 0.2, etc., we randomly selected five scenarios. Figure S3 of supplementary material available at *Biostatistics* online shows the power of D vs. D_{ratio}. The statistics were comparable in power, although D_{ratio} seemed modestly less powerful in the mid-range.

3.2 *Additional investigations of type I error rate control*

We assessed the performance of our analytic approximations using annotated pathways for two datasets: (i) the saliva data described earlier and (ii) a breast cancer dataset from Miller *and others* (2005), (GSE3493 Affymetrix U133A, where $n = 251$, $m = 22\,215$). Martingale residuals for disease-free survival were available as a continuous **x** for $n = 236$, and mutational status of p53 ($-/+$), a well-known tumor marker, as a dichotomous **x** ($n_0 = 198$, $n_1 = 56$). We selected a KEGG pathway for the salivary dataset and gene ontology (GO) pathways for the breast cancer dataset.

Tables 1 and 2 display the type I error rate results for U, V, and D for the illustrative pathways. For dichotomous **x** in the salivary gland dataset, the true thresholds for type I error rate control were established by exhaustive permutation (48 620), and by 10^8 random permutations for other table entries. We emphasize that this large number of permutations for all pathways is impractical for standard use, and is used here to establish the accuracy of our approach. To cover the range of useful pathway testing thresholds, we display ratio$_\alpha$ = (true Type I error rate)/α, for α ranging from 10^{-1} to 10^{-6}. The proposed approximations are reasonably accurate.

Table 2. *Performance of the p-value approximations for GO:0000184* (44 *genes*), *with continuous breast cancer disease-free survival* ($n = 236$) *and dichotomous p53 status* ($n_0 = 198, n_1 = 56$)

Threshold	Proposed approximation					
	Continuous **x**			Dichotomous **x**		
α	U	V	D	U	V	D
10^{-1}	1.00	1.01	1.01	1.00	1.01	1.01
10^{-2}	1.00	0.99	0.96	1.00	0.99	0.96
10^{-3}	1.01	0.96	0.92	1.01	0.98	0.93
10^{-4}	1.07	0.97	0.82	1.07	1.00	0.92
10^{-5}	1.27	0.85	0.84	1.24	1.32	0.78
10^{-6}	1.52	0.45	0.76	1.50	0.91	0.71

Entries in the table are ratio_α.

As V can be expressed as a quadratic form, there are competing analytic approximations that have been considered, not necessarily in a genomics context. Two popular approaches are based on (i) a scaled central χ^2 distribution (Duchesne and Micheaux, 2010), for which the first two moments are typically used for fitting, and (ii) a non-central χ^2 distribution (Liu *and others*, 2009), which is used in the globaltest with finite sample corrections. The globaltest results are shown in Table S1 of supplementary material available at *Biostatistics* online for the salivary gland dataset, for which values less than $\alpha = 10^{-5}$ begin to show inaccuracy. For the dataset under dichotomous **x**, the smallest globaltest p-values are less than $\alpha = 10^{-5}$, even though there are only 24 310 unique permuted V values. Using approximate χ^2 results that are based on the naive moments are extremely poor, with true false positive rates differing by 1–2 orders of magnitude from that intended, and are not shown.

To demonstrate the effectiveness of the covariate adjustment, we used KEGG:00510 (92 genes) for the salivary data. We consider the scenario that provides challenges for proper covariate adjustment: (i) a small sample size; (ii) two covariates, one which is correlated with both $\mathbf{Y}_{\text{path}}$ and **x**. Recall that, in earlier results, the data (both expression and clinical/treatment variable) could be considered fixed, and it was possible for each pathway to create a single permutation distribution to which the analytic approximations could be compared. If the type I error rates are always controlled, conditioned on the observed data, then it follows that the type I error rate is also controlled unconditionally. Here, however, covariates disrupt the interpretation of the permutation distribution, and to establish control of the type I error rate, we condition on $\mathbf{Y}_{\text{path}}$, but generate random realizations of **x** and **Z**. For this example, covariate z_1 was generated as $0.2 \times \mathbf{y}'$ of $\mathbf{Y}_{\text{path}}$, plus an $N(0, 1)$ error term. A dichotomous **x** was generated using the logistic model with $\text{logit}(y) = z_1$, and rejection sampling to ensure that $n_0 = 9, n_1 = 9$. Covariate z_2 was generated as $N(0, 1)$. After applying the covariate residualization and proper effective sample size, qqplot results for 10 000 data simulations show good performance of the resulting p-values using our approximations for U_z and V_z (Figure S4 of supplementary material available at *Biostatistics* online). Failing to adjust for covariates, or incorrectly failing to account for the reduced rank of the weighted beta approximation, both produce highly inaccurate p-values (Figure S5 of supplementary material available at *Biostatistics* online). Similar results hold for continuous **x**, for which an example is shown in Figure S6 of supplementary material available at *Biostatistics* online.

To further illustrate the performance and efficiency of our methods, we analyzed the breast cancer data with a total of $K = 6701$ GO pathways, using 10^8 random permutations for the p53 status phenotype, and also using *safeExpress*. *safeExpress* is several orders of magnitude faster than random permutation for individual pathways. Direct permutation required tens of thousands of total CPU hours on a computing cluster, while *safeExpress* took at most a few hours on a standard PC (depending on the grid density; see Table S1

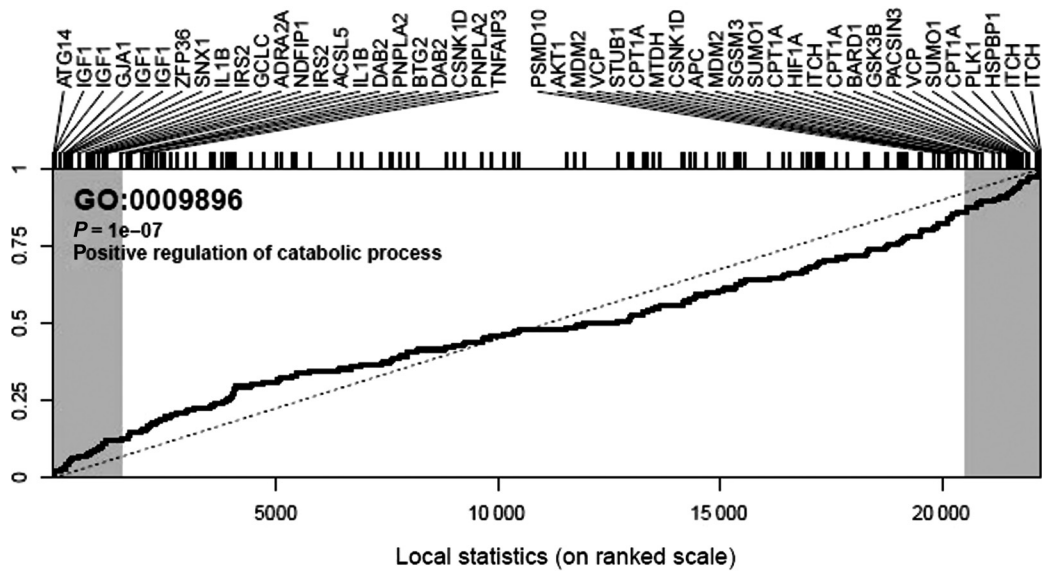

Fig. 2. Safe-plot of ranks of local S statistics, using disease-free survival for the breast cancer data, GO: 0009896. The plot shows an empirical cumulative distribution function of the ranks of the local score statistics S_i within the pathway, interspersed among the ranked local statistics of genes in the complement. Genes in the pathway tend to have extreme positive or negative score statistics, representing poor or good prognosis when expression is high, respectively.

of supplementary material available at *Biostatistics* online). Of course, 10^8 permutations are unnecessary for many pathways, as many pathways can be declared non-significant after only a few permutations. We show in Appendix E of supplementary material available at *Biostatistics* online that an adaptive permutation procedure, designed to avoid unnecessary permutations, still requires 40+ times as much computation as *safeExpress* in order to effectively control false positives. In addition, *safeExpress* is designed to take advantage of the R *multicore* package to gain further efficiency from parallel processing.

Using *safeExpress*, we found several highly significant GO pathways for the various **x** vectors: (i) p53 mutation status in all patients; (ii) p53 mutation status for estrogen-receptor positive (ER+) tumors only ($n = 213$); (iii) p53 mutation status in all patients, with covariate adjustment for ER status; and (iv) the continuous martingale residuals for disease-free survival. The most significant pathways, with Benjamini–Hochberg q-values computed for each of U, V, and D are provided as supplementary files. Here, we highlight just a few results from (iii) to (iv), with potential supporting literature. Results for (iii) included GO:0000778, Condensed Nuclear Chromosome Kinetochore ($p = 5.2 \times 10^{-15}$ for U) (Chi *and others*, 2009), and GO:0045767, Regulation of Anti-apoptosis ($p = 8.5 \times 10^{-10}$ for D). For (iv) (disease-free survival), fewer pathways were significant after multiple test correction, but included GO:0051087, Chaperone Binding ($p = 3 \times 10^{-6}$ for U) (Marx *and others*, 2007), and GO:0009896 Positive Regulation of Catabolic Process ($p = 1 \times 10^{-7}$ for D) (Wallace *and others*, 2000), a pathway that includes *IGF-1*, a marker widely studied for its role in breast and other cancers (Chong *and others*, 2011). For the last finding, a "safe-plot" (Barry *and others*, 2005) is shown in Figure 2, showing the ranks of local statistics within the pathway.

Figure 3 compares the $-\log_{10}(p)$-values from the 10^8 random permutations vs. the *safeExpress* approximations for U, V, and D across all 6701 pathways for testing p53 mutation status **x**. The agreement is close, and variation for large $-\log_{10}(p)$ is largely due to sampling variation from the random permutations,

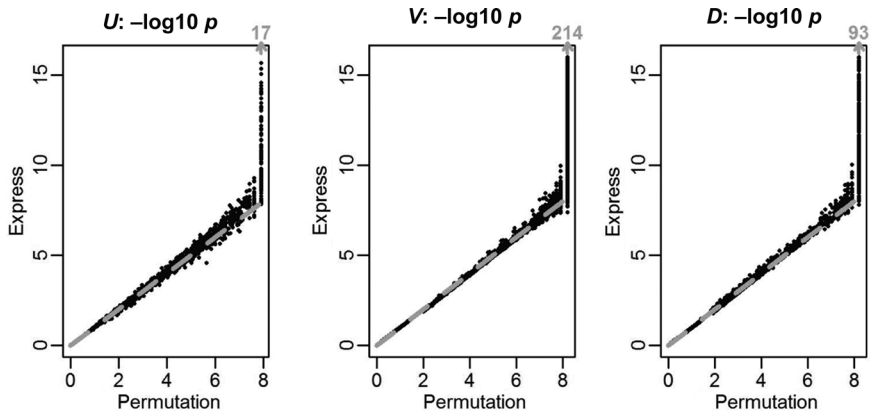

Fig. 3. For the breast cancer data with a p53 dichotomous outcome, comparison between *safeExpress* and 10^8 random permutations on the $-\log 10(p)$ scale. The numbers of pathways exceeding the y-axis limit are shown in gray.

not from approximation bias. Note that, with 10^8 permutations, standard permutation-based p-values will not be less than 10^{-8}, even when the data indicate that the exact p-value is much smaller. In contrast, *safeExpress* can identify and rank pathways that are much more highly significant.

4. Discussion

We have demonstrated that an analytic approach can provide approximations to permutation p-values for gene pathway testing, of high enough accuracy to substitute for permutation. Although confined to statistics that involve linear operations, our results include several novel techniques that are likely useful in other contexts. The corrected beta r^2 approximation applies generally to simple regression problems, and the weighted beta approximation applies to sums of squared score statistics, which are widely used in ensemble hypothesis testing. We are not aware that accurate analytic approximations to mean differences of squared score sums (D, in our context) have been previously proposed.

For our speed comparisons, the random permutation approach was applied simultaneously to the entire expression matrix, and then summarized within each category. This an efficient approach when computing results for numerous categories. For a single category for r_U, it is feasible to compute a large number of random, though not entirely independent, permutations by computing cross-products of matrices consisting of random permutations of $\mathbf{x}$ and $\mathbf{y}'$. Permutations of V using matrices of permuted $\mathbf{x}$ are also possible, though slower and with considerable RAM requirements. Using this approach for D (which uses all genes) quickly becomes unwieldy. As the number of principal components never exceeds $n - 1$, another possibility is to permute the PCs. However, the simple nature of our analytic approximations, and the relative ease of embedding the approximations into a software implementation, make *safeExpress* especially attractive.

Extensions to our work include potential applications in SNP-set testing, sequence-based association analysis, and other 'omics applications involving sets of correlated statistics. For some of those applications, the accuracy of the approximations must be demonstrated to even more stringent thresholds to account for an even larger number of tests.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Dr Wei Pan for pointing out the connection to the globaltest statistic. Portions of this work were conducted while Dr Barry was appointed at Duke University. *Conflict of Interest:* None declared.

Funding

This work was supported in part by the Gillings Statistical Genomics Innovation Lab, EPA RD83382501, NCI P01CA142538, and NIEHS P30ES010126 and P42ES005948.

References

Barry, W. T., Nobel, A. B. and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943–1949.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics* **2**, 286–315.

Buzkova, P., Lumley, T. and Rice, K. (2011). Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions. *Annals of Human Genetics* **75**, 36–45.

Chi, Y. H., Ward, J. M., Cheng, L. I., Yasunaga, J. and Jeang, K. T. (2009). Spindle assembly checkpoint and p53 deficiencies cooperate for tumorigenesis in mice. *International Journal of Cancer* **124**, 1483–1489.

Chong, K, Subramanian, A., Sharma, A. and Mokbel, K. (2011). Measuring IGF-1, ER-α and EGFR expression can predict tamoxifen-resistance in ER-positive breast cancer. *Anticancer Research* **31**, 23–32.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **8**, 242.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2009). Gene-set analysis and reduction. *Briefings in Bioinformatics* **10**, 24–34.

Duchesne, P. and Micheaux, P. L. D. (2010). Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics and Data Analysis* **54**, 858–862.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Human Genetics* **8**, 376–386.

Fridley, B. L., Jenkins, G. D. and Biernacka, J. M. (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS ONE* **5**, e12693.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I. and Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics* **11**, 574.

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

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

Goeman, J. J. and van Houwelingen, H. C. (2011). Testing against a high-dimensional alternative in the generalized linear model asymptotic type I error control. *Biometrika* **98**, 381–390.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21**, 1950–1957.

Harris, P. and Peers, H. W. (1980). The local power of the efficient scores test statistic. *Biometrika* **67**, 525–529.

Kennedy, P. E. and Cade, B. S. (1996). Randomization tests for multiple regression. *Communications in Statistics Simulation* **25**, 923–936.

Knijnenburg, T. A., Wessets, L. F. A., Reinders, M. J. T. and Shmulevich, H. (2009). Fewer permutations, more accurate P-values. *Bioinformatics* **25**, 161–168.

Lee, S., Wright, F. A. and Zou, F. (2011). Control of population stratification by correlation-selected principal components. *Biometrics* **67**, 967–974.

Liu, Z. J. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics* **87**, 139–145.

Liu, Q., Dinu, I., Adewale, A. J., Potter, J. D. and Yasui, Y. (2007). Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* **8**, 431.

Liu, H., Tang, Y. and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis* **53**, 853–856.

Marx, C., Yau, C., Banwait, S., Zhou, Y., Scott, G. K., Hann, B., Park, J. W. and Benz, C. C. (2007). Proteasome-regulated ERBB2 and estrogen receptor pathways in breast cancer. *Molecular Pharmacology* **71**, 1525–1534.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T. *and others*. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13550–13555.

Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* **33**, 497–507.

Schaid, D. J. and Sommer, S. S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents. *American Journal of Human Genetics* **55**, 402–409.

Subramanian, A. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550.

Therneau, T. M. and Grambsch, P. M. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.

Virtaneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de la Chapelle, A. and Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proceedings of the National Academy of Sciences* **98**, 1124–1129.

Wallace, H. M., Duthie, J., Duthie, J., Evans, D. M., Lamond, S., Nicoll, K. M. and Heys, S. D. (2000). Alterations in polyamine catabolic enzymes in human breast cancer tissue. *Clinical Cancer Research* **6**, 3657–3661.

[*Received September 18, 2012; revised December 20, 2012; accepted for publication January 19, 2013*]