University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2012 *Paper* 24

Empirical Pathway Analysis, without Permutation

Yi-Hui Zhou[*] William T. Barry[†]
Fred A. Wright[‡]

[*] University of North Carolina at Chapel Hill, yihuiz@live.unc.edu
[†] Duke University, bill.barry@duke.edu
[‡] University of North Carolina at Chapel Hill, fred_wright@unc.edu

Empirical Pathway Analysis, without Permutation

Yi-Hui Zhou, William T. Barry, and Fred A. Wright

Abstract

Resampling-based expression pathway analysis techniques have been shown to preserve type I error, in contrast to simple gene-list approaches which implicitly assume independence of genes in ranked lists. However, resampling is intensive in computation time and memory requirements. We describe highly accurate analytic approximations to permutations of score statistics, including novel approaches for Pearson correlation and summed score statistics, that have good performance for even relatively small sample sizes. In addition, the approach provides insight into the permutation approach itself, and summary properties of the data that largely determine the behavior of the statistics. Within the framework of the SAFE pathway analysis procedure, our approach preserves the essence of permutation analysis, but with greatly reduced computation. Extensions to include covariates are described, and we test the performance of our procedures using simulations based on real datasets of modest size.

Keywords: gene set analysis; permutation; hypothesis testing.

Empirical Pathway Analysis, Without Permutation

Yi-Hui Zhou, William T. Barry, Fred A. Wright

Resampling-based expression pathway analysis techniques have been shown to preserve type I error, in contrast to simple gene-list approaches which implicitly assume independence of genes in ranked lists. However, resampling is intensive in computation time and memory requirements. We describe highly accurate analytic approximations to permutations of score statistics, including novel approaches for Pearson correlation and summed score statistics, that have good performance for even relatively small sample sizes. In addition, the approach provides insight into the permutation approach itself, and summary properties of the data that largely determine the behavior of the statistics. Within the framework of the SAFE pathway analysis procedure, our approach preserves the essence of permutation analysis, but with greatly reduced computation. Extensions to include covariates are described, and we test the performance of our procedures using simulations based on real datasets of modest size.

Keywords: gene set analysis; permutation; hypothesis testing.

1 Introduction

A basic approach to gene expression analysis involves the detection of genes significantly differentially expressed among treatment conditions, or more generally exhibiting association between expression and a clinical outcome or experimental design variable (hereafter called the *response*). This kind of analysis focuses on individual genes. However, researchers are often interested in association of the response with sets of genes of related biological function, either to increase power or to provide a more parsimonious, pathway-based view of the results. A large number of methods and software for gene-set analysis have been proposed (Dinu *et al.* (2009)), and can be divided into approaches that implicitly assume uncorrelated expression data vs. those that acknowledge such correlation (Barry *et al.* (2008)). As described in Gatti *et al.* (2010), approaches that acknowledge correlation via permutation have vastly superior type I error control compared to methods that assume no correlation structure, and include GSEA (Subramanian (2005) and Mootha *et al.* (2003)), SAFE (Barry *et al.* (2005)), and additional methods (Pollard & van der Laan (2005)). The *globaltest* of Goeman *et al.* (2004) offers a non-resampling approach, in which the correlations structures are parametrically modeled, in turn producing estimates of the mean and variability for a particular score-related statistic.

In addition to its proper handling of correlation structures, the SAFE methodology offers a useful distinction between *local* statistics, which measure the association between individual genes and the response, and *global* statistics, which are aggregations of local statistics, and used to describe the overall association of a gene set with the response. The use of sample permutation-based gene set analysis dates to Virtaneva *et al.* (2001), in which the response is permuted relative to the expression data, with the entire dataset analyzed for each permutation. The appeal of permutation is that it enables conditioning on the existing gene expression structure without explicit modeling, while reflecting the downstream effect of that structure on statistical analysis. As described in detail in Gatti *et al.* (2010), proper handling of correlation among genes in a set is essential for error control – otherwise the coincident appearance of highly correlated genes among a significant gene list appears spuriously significant. In addition, permutation enables handling of correlation across different gene sets. This latter issue is of lesser importance, as false discovery techniques are often used for error control for multiple gene sets, and are typically robust to positive correlation of tests (Kim & van de Wiel (2008)).

A downside to permutation approaches is that they are highly computationally intensive, keeping in mind that all genes and all categories are examined for each permutation. The SAFE and GSEA procedures compute and store the entire resampled matrices of global statistics for careful error control. 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 *et al.* (2009)), and effective control of the false discovery rate (FDR) or family-wise error across categories may not be possible, as the empirical p-values have a minimum 1/(number of permutations).

As an alternative viewpoint, we note that correlation among local test statistics is induced by correlation among genes, as discussed in some detail in Barry *et al.* (2008). It is thus worth exploring whether the null distributions of certain local and global statistics are amenable to parametric approximation, using empirical estimates of correlation structure. In other words, is it possible to perform gene set testing based conceptually on sample permutation, but without actually permuting at all? We explore that possibility in this paper, and provide comparisons to less accurate standard parametric approximations. For a proposed global statistic comparing each pathway to its complement, we are not aware that competing parametric approximations have been proposed, due to the perceived difficulty in handling the correlation structures.

2 Methods

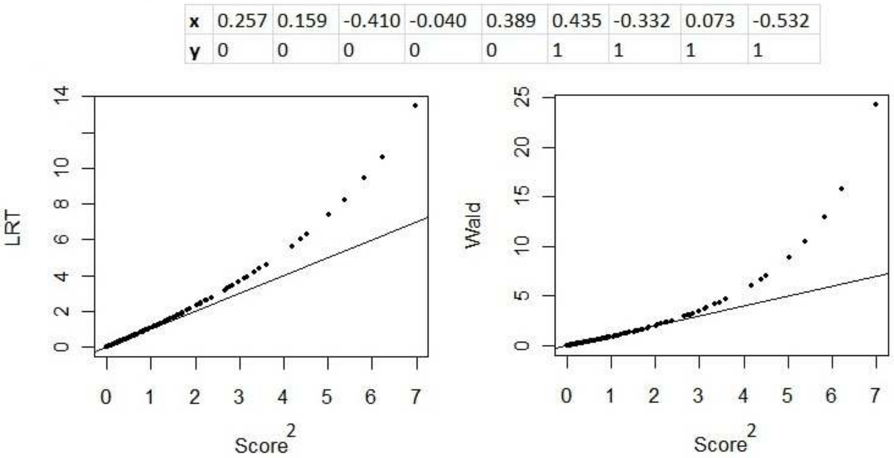
2.1 Notation

We suppose that each sample $j = 1, 2, ..., n$ is associated with a response value y_j. We use $\mathbf{y}$ to denote the response vector, which might be discrete or continuous, and we later describe approximate procedures to handle censored time-to-event data. We use g_{ij} to denote the gene expression level for the ith gene ($i = 1, 2, ..., m$) and jth sample. Let $\mathbf{X}$ be the $m \times n$ normalized gene expression matrix, with $x_{ij} = \frac{g_{ij} - \overline{g_{i.}}}{\sqrt{\sum_{j=1}^{n}(g_{ij} - \overline{g_{i.}})^2/n}}$, where $\overline{g_{i.}} = \sum_j g_{ij}/n$. This normalization produces $\sum_{j=1}^{n} x_{ij} = 0$ and $\sum_{j=1}^{n} x_{ij}^2 = 1$, providing convenient simplifications for later development.

An important choice of local statistic to represent the relationship between the ith gene and the response is the score statistic, which for linear regression and a variety of generalized linear models can be expressed as (Schaid & Sommer (1994))

$$S_i = \frac{\mathbf{x}_{i.}^T \mathbf{y}}{\sqrt{\Sigma_{j=1}^{n}(y_j - \overline{y})^2/n}} = \frac{\sum_j x_{ij} y_j}{s_y \sqrt{(n-1)/n}}, \tag{1}$$

where s_y is the sample standard deviation of the $\mathbf{y}$ values. Our notation thus far largely follows that of Lee *et al.* (2011) (except for a square root difference in S_i), 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 (Buse (1982)), although in certain small-sample settings other statistics may have slight advantages (Harris & Peers (1980)). One source of confusion arises from improper control of type I error, which can produce the illusion of reduced power for a statistic that otherwise is good at detecting departures from the null. This paper is concerned with permutation testing, and in such settings the various "standard" statistics may be one-to-one across permutations (Gatti *et al.* (2009)), and therefore have equivalent



x	0.257	0.159	-0.410	-0.040	0.389	0.435	-0.332	0.073	-0.532
y	0	0	0	0	0	1	1	1	1

Figure 1: For the data shown, the score statistics, Wald statistics and likelihood ratios are permutation-equivalent. The 9! permutations collapse into $\binom{9}{4} = 126$ unique values, because $\mathbf{y}$ is dichotomous.

power. Figure 1 illustrates this simple concept for an example of a single simulated gene and dichotomous response for $n = 9$, for which the squared score statistics, logistic regression Wald statistics, and likelihood ratios are monotone with each other.

Hereafter we use the term *category* to refer to the set of genes under study, which is more generic than the evocative term *pathway.* In choosing a global statistic, a key consideration is whether the direction of expression-response association (positive or negative) is expected to be the same for the associated genes. Another consideration is whether the appropriate null distribution should be for the "self-contained" hypothesis (that no gene in the category exhibits differential expression) or the "competitive" hypothesis that the degree of differential expression is the same within the category as in the remaining complementary set of genes. The self-contained/competitive nomenclature has been laid out by Goeman & Buhlmann (2007), and a variety of accompanying global statistics are described in detail in Barry *et al.* (2008). Within the score statistic framework, we propose global statistics based on simple sums and differences, described in further detail below.

2.2 Self-contained testing

Under the complete null hypothesis that no gene shows differential expression, it is reasonable to focus on one category at a time, ignoring the evidence from the remainder of genes. We will use $\{cat\}$ to denote a category containing m_{cat} genes. We (Barry *et al.* (2008)) have criticized self-contained testing, as it fails to account for gene effects that are not specific to a category. Nonetheless, there are a number of situations in which self-contained testing may be reasonable. These may 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 category is tested, and where any evidence of significance enrichment is of interest.

2.2.1 The directional statistic U

The global statistic $U = \sum_{i \in cat} S_i$ is perhaps the most straightforward directional statistic, in the sense that it is sensitive to expression-response associations that are in the same direction. Nonetheless, testing for this statistic will generally be two-sided. A normal approximation to U might seem appealing, as summation is performed both over n (within S_i) and m_{cat}, and it might seem that the central limit theorem should apply especially well for large m_{cat}. However, the distribution of U is fundamentally limited by n, as shown by rewriting

$$U = \frac{\sum_{i \in cat} \sum_j x_{ij} y_j}{s_y \sqrt{(n-1)/n}} \propto \sum_j \left(\sum_{i \in cat} x_{ij} \right) y_j = \sum_j x'_j y_j,$$

for $x'_j = \sum_{i \in cat} x_{ij}$ (with the vector of these values denoted $\mathbf{x}'$). In this formulation, it is simple to show (Appendix A1) that under permutation U is one-to-one with the Pearson correlation r_U between $\mathbf{x}'$ and permutations of $\mathbf{y}$. This immediately suggests the standard test statistic for association, $t_U = r_U \sqrt{(n-2)/(1-r_U^2)}$ (Jobson *et al.* (1991)). It is known that t_U follows a t_{n-2} density under the null of independent $\mathbf{x}'$ and $\mathbf{y}$, provided at least one of the two variables is normal. Hereafter, we will refer to the corresponding density of r_U (a signed square root of a beta density) as *standard* r. For r_U under permutation, the variance is exactly $1/(n-1)$, and the standard r density may be used as an approximation, but accuracy will suffer if the empirical cdfs of both $\mathbf{x}'$ and $\mathbf{y}$ are far from normal (see Appendix A1). For continuous $\mathbf{y}$, the fit of standard r is relatively accurate, especially if the empirical distribution of $\mathbf{y}$ is nearly normal. Although the sample correlation has been extensively studied, analytic results for non-normal data have been mainly limited to special instances such as a mixture of bivariate normals (Srivastava & Lee (1984)).

An alternate approach is to use a saddlepoint approximation (Daniels (1954)), which has been proposed to avoid permutation for two-sample testing (Robinson (1982)), and is equivalent to testing r_U for dichotomous $\mathbf{y}$ (Appendix A2). The tail accuracy of two-sample saddlepoint approximations (equivalent to our r_U) have often been compared to that of a normal approximation (e.g. Robinson (1982), Abd-Elfattah (2009)) although comparison to the standard r density is more appropriate. Thus the utility of the saddlepoint in this context is difficult to judge. Moreover, the saddlepoint is not available for continuous $\mathbf{y}$.

As a simple example where standard r can fail, we consider the salivary gland expression data (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 to $n_0 = 9$ controls. For the 92 probe sets in KEGG:00510, N-Glyean, Biosynthesis, Figure 2 shows the distribution of r_U over the $\binom{18}{9} = 48,620$ unique permutations, overlaid with the standard r. In Appendix B4 we describe an alternate approximating density for r_U, which is a rescaled equal mixture of two standard r densities, separated by an equal offset about zero. The new approximating density is unskewed, and we may use its kurtosis to determine the offset value. The new kurtosis-corrected density provides a superior approximation to the permutation tails(Figure 2). Importantly, for dichotomous $\mathbf{y}$ the kurtosis of r_U can be determined directly from the kurtosis of $\mathbf{x}'$ (treating the observed $\mathbf{x}'$ as a population, Appendix A3), and thus no actual

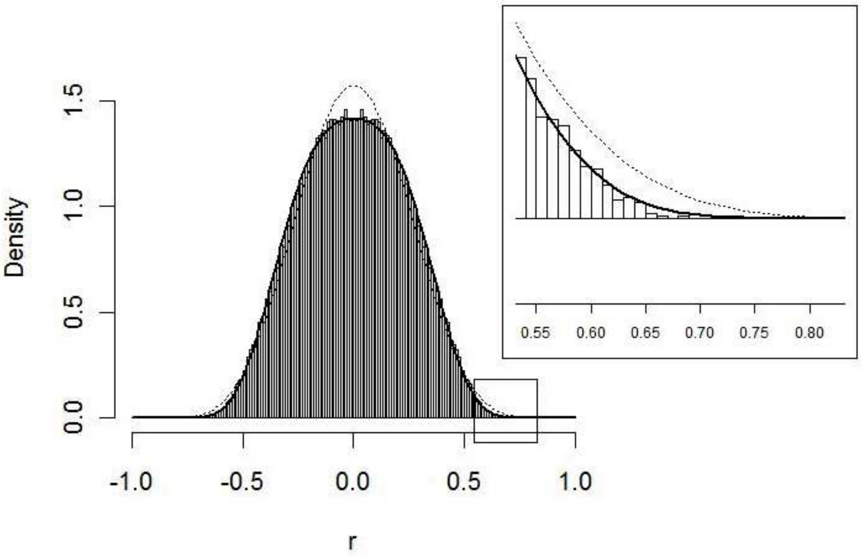


Figure 2: Comparison of fitting r_U by the standard r (dashed curve) and kurtosis-corrected mixture (solid curve) approximations, for exhaustive permutations of the salivary data, KEGG:00510 .

permutation is required to fit the kurtosis-corrected density.

Finally, although the kurtosis-corrected density already provides accurate fits to r_U under dichotomous $\mathbf{y}$, the fit can be further improved in the tails by considering the permutations which achieve maximum and minimum r_U, which can be easily obtained from the order statistics for $\mathbf{x}'$. The approximating values at these extremes can then be compared to the true directional p-values, known to be $1/\binom{n}{n_1}$. Briefly, the discrepancy between the approximating and true extreme p-values is expressed in terms of a quantile for the minimum p-value among $\binom{n}{n_1}$ uniformly distributed p-values. This quantile is then used to adjust all of the kurtosis-corrected estimated p-values that achieve some modest significant level (e.g., $p < 0.05$).

2.2.2 Examples, and the required accuracy for an effective approximation

To illustrate these concepts, Figure 3 shows the performance of the various approaches for exhaustive permutation in the two-sample (dichotomous $\mathbf{y}$) setting, using an illustrative KEGG category for the saliva dataset. Treating the p-values from exhaustive permutation as a gold standard, the offset-mixture r approximation has generally good performance, especially with the quantile correction. Its performance is similar to that of the iterative saddlepoint method, even when the moment-generating function for the saddlepoint is determined exactly from the exhaustive permutations (Appendix A2).

In practice, a random sample of permutations is used, and this number is typically 10,000 or fewer, due to computational constraints. When testing numerous categories, empirical p-values in the range of 10^{-4} or smaller may be necessary to achieve significance. Thus, although random permutation-based p-values are unbiased, the relative variability remains problematic. In contrast, the analytic approximations described here may be biased, but for fixed data are not subject to sampling variability.

A natural question arises: when should an analytic approximation to the true permutation p-value be considered superior to that derived from π random permutations? We adopt the following heuristic argument, based on mean-squared error (MSE). Suppose α is the intended type I error, while $f\alpha$ is the true error associated with the analytic approximation. Thus, f is an error ratio, and values near 1.0 are desired. For small α, the MSE (here equal to the variance) of permutation rejection proportions is $\approx \alpha/\pi$, while the MSE (bias^2) of the analytic approximation is $\alpha^2(1-f)^2$. When $\alpha = 1/\pi$ (which occurs for $\alpha = 10^{-4}$ and $\pi = 10,000$), a simple computation shows that the MSE for the analytic approximation is more favorable as long as $f \in (0, 2)$. In practice, to guard against excessive conservativeness, we will consider the analytic approximation to be superior if $f \in (0.5, 2)$.

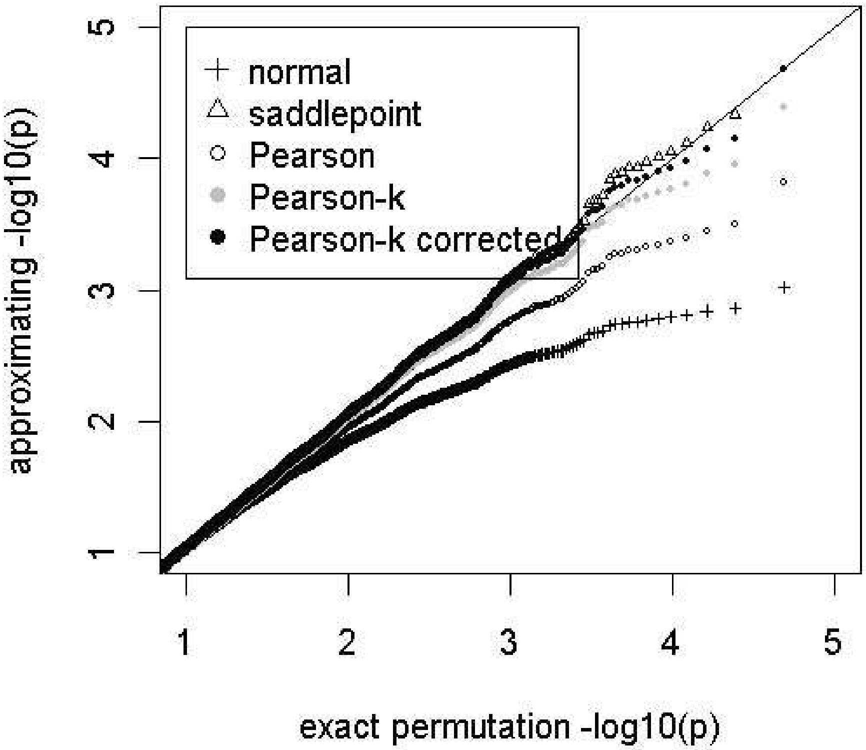


Figure 3: Right tail p-values for U, the KEGG:00510 category and the salivary data ($n_0 = n_1 = 9$).

2.2.3 The non-directional statistic V

One of the simplest non-directional global statistics to detect departures from the self-contained null is a summation of the squared score statistics in the category,

$$V = \sum_{i \in cat} S_i^2 \ . \tag{2}$$

This statistic is now widely used in gene/SNP-set testing for association analysis (Liu (2010)). Moreover, for the expression context it is equivalent to the *globaltest* (Goeman *et al.* (2004)) when the expression data have been scaled by gene (Pan (2009)). Goeman (Goeman *et al.* (2004)) have argued that V is optimal when aggregating for small effect sizes, similar to arguments by Newton *et al.* (2007). The standard approach to estimating the distribution of V is to use quadratic form results and moment-matching (Liu *et al.* (2009)) to obtain approximations based on non-central (Liu *et al.* (2009)), or scaled (Duchesne & Micheaux (2010)) χ^2. The "naive" moment estimates here are $E(V) = m_{cat}$ (following from a χ_1^2 approximation for S_i^2), and var(V) = 2trace($R^T R$), where $R = \mathbf{X}_{cat}^T \mathbf{X}_{cat}$ is the correlation matrix for the genes in $\{cat\}$. However, for small to moderate sample sizes, the departure from these moments has a noticeable impact on approximation accuracy.

Before describing the moments in detail, we first highlight an alternate formulation, described in Lee *et al.* (2011):

$$V = n \sum_{j=1}^{n} \lambda_j r_j^2, \tag{3}$$

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

Equation (3) shows precisely how the rank of $\mathbf{X}_{cat}$ affects V. If $m_{cat} < n - 1$, some of the λ_j will be zero, and even for large m_{cat}, the number of contributing terms cannot exceed $n-1$. There is no need to otherwise account for the rank of $\mathbf{X}_{cat}$. Note that the orthogonality of the first $n-1$ PCs implies that $\sum_j r_j^2 = 1$, while for constant V, (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 chisquare-based approximations to V can fail in the extreme tail, even if the moments are specified correctly. Finally, we emphasize that the correspondence between (2) and (3) is exact, so that permutations of V could be equivalently obtained by instead recomputing (3) across permutations of principal components.

Here we use (3) to motivate an alternative analytic approximation to the permutation distribution of V. The standard r approximation implies that each $r_j^2 \sim Beta(1/2, (n-2)/2)$ (Fisher (1938)), which has the same variance as the earlier permutation result $\text{var}(r_j^2) = E(r_j^2) = 1/(n-1)$, noting that any of the PCs can serve the same role played by $\mathbf{x}'$ in the earlier subsection. The naive moments can be derived if one assumes (incorrectly) that $E(r_j^2) = 1/n$, and that the r_j^2 terms are uncorrelated. The $\{r_j^2\}$ are actually somewhat negatively correlated, as shown in a general "correlation of squared correlation" result in Appendix B2, for predictors that are not necessarily orthogonal. For the orthogonal PCs, we use a multivariate normality assumption for both $\mathbf{y}$ and $\mathbf{P}$ to imply that the $\{r_j^2\}$ follow a correlated joint beta density, conceptually related to the multiple correlation coefficient sampling distribution (Fisher (1928)). The joint density is derived in Appendix B1, 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 approximation $r_1^2 \sim Beta(1/2, (n-2)/2)$, and show that for any subset $\Omega \subset \{1, ..., n-1\}$ which doesn't contain k,

$$\frac{r_k^2}{1 - \sum_{j\in\Omega} r_j^2} \sim Beta(1/2, (n - |\Omega| - 2)/2).$$

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 $Beta(1/2, (n-1-j)/2)$. Thus V can be approximated as a sum of weighted correlated beta variates (hereafter referred to as the *weighted beta*). The approximation reflects correlation structure that may be non-trivial for moderate sample size, and 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 the numeric approximation below.

Although the weighted beta approximation is accurate, computing tail probabilities is computationally intensive. Thus we use a combination of numeric integration for the initial terms $\lambda_1 r_1^2 + \lambda_2 r_2^2$, and a shifted gamma approximation for the remaining terms in (3). The shifted gamma distribution is an ordinary gamma with an additional location parameter, and the three parameters are computed using moment-matching from the eigenvalues. The shifted gamma density in fact provides a reasonable fit to the entire V distribution, but tends to have overly heavy tails. In our implementation, for extreme V the first two PCs dominate, so that inaccuracies caused by the gamma fit are minimized.

2.3 Competitive testing

Competitive global test statistics contrast the local statistics within each category vs. those of the complementary set of genes (Goeman & Buhlmann (2007)). For datasets in which a large number of genes are associated with $\mathbf{y}$, this approach enables the researcher to focus on results that are truly related to the category, 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 correlation structure in each category vs. its complement (Barry *et al.* (2008)).

For aggregations of directional local statistics, a straightforward competitive global statistic is $\frac{U_{cat}}{m_{cat}} - \frac{U_{comp}}{m_{comp}}$, where $\{comp\}$ designates the complementary set of genes not in $\{cat\}$.

However, for array studies we must consider the role of data normalization, and we note that $U_{all} = U_{cat} + U_{comp}$. 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_{cat} and U_{comp} will have correlation -1 over permutations. Thus the construction of a competitive global statistic would be redundant. In other words, U_{cat} is essentially *already* a competitive statistic, as these normalization procedures force the average effect across all genes to be zero. 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_{cat} should not be further modified in an attempt to make it "competitive."

2.3.1 The competitive statistic D

The natural competitive global statistic based on sums of squared score statistics is

$$D = V_{cat}/m_{cat} - V_{comp}/m_{comp}, \tag{4}$$

which is not subject to the normalization issues described above. As m_{comp} is typically much larger than m_{cat}, one might expect that V_{comp}/m_{comp} could be treated as nearly constant. In fact, for many datasets, the gene-gene correlation structures are strong enough that variation in V_{comp}/m_{comp} is non-negligible. Moreover, V_{cat} and V_{comp} are correlated. Although V_{comp} may also be approximated by the beta mixture approach, there is no reason to expect that the difference D can be easily modeled using any of the procedures above, and we are not aware of any reported approximations that are applicable. The earlier correlation result from (2.2.3) can be used to obtain the correlation of V_{cat} and V_{comp} through the correlations of their respective principal components (involving at most $n \times n$ terms), thus providing the permutation variance of D. However, obtaining higher moments through this approach appears unwieldy for fast computation.

The primary barrier to approximating the distribution of D is that the form of (2) cannot be negative, involving sums of squared terms, while its equivalent (3) is critical to the beta-mixture approach. As we show in detail in Appendix C, this obstacle is neatly overcome by adding weights w_i to the score statistics $\{S_i^2\}$, where the weights for $i \in \{comp\}$ are *imaginary*, resulting in negative squared terms. Formally, we create a new weighted matrix $\mathbf{A} = \mathbf{WX}$, where $\mathbf{W}$ is the diagonal matrix with terms $\{w_i\}$. Here a new equivalent relation holds,

$$D = \sum_{i=1}^{m} w_i^2 S_i^2 = n \sum_{j=1}^{n} \gamma_j c_j^2, \tag{5}$$

where $\{\gamma_j\}$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$, and c_j^2 are the observed squared correlations between the corresponding eigenvectors and $\mathbf{y}$. Our use of imaginary terms is a useful device, but $\mathbf{A}^T\mathbf{A}$ is a real matrix, and so no complex algebra is required for the decomposition. Moreover, computation is greatly simplified by computing $\mathbf{X}^T\mathbf{X}$ once, and then computing $\mathbf{A}^T\mathbf{A} = \mathbf{X}^T\mathbf{X} - \mathbf{X}_{cat}{}^T\mathbf{X}_{cat}$ for each category. Equation (5) 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).

2.4 Inclusion of covariates

Suppose $\mathbf{z}_1$, $\mathbf{z}_2$, ...,$\mathbf{z}_p$ are a set of n-vector covariates, any of which may be correlated with $\mathbf{X}$ or $\mathbf{y}$, 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 correlation structures across genes, for which permutation is attractive. The proper handling of covariates is a challenge in the permutation setting, however, as standard permutation forces the investigator to permute the covariates relative to either $\mathbf{X}$ or $\mathbf{y}$. Such an approach is inappropriate if a covariate is correlated with both $\mathbf{X}$ and $\mathbf{y}$. Several permutation approaches in the presence of covariates are described in Good (2000) for linear regression. All of these approaches involve fitting a full regression model, including both covariates and effects of interest, followed by construction of new data in which permuted residuals from the full model are added to fitted values from a reduced model in which the predictor of interest has been excluded. This approach is difficult to model analytically, and we consider a simpler approach described below.

We begin by noting that, in the presence of covariates, the natural analog to our score statistic is $S_{i,z} = \sum_{j=1}^{m_{cat}} x_{ij,z} y_{j,z} / \sqrt{\sum_j \frac{(y_{j,z} - \overline{y}_z)^2}{n}}$, where $\mathbf{x}_{i.,z}$ and $\mathbf{y}_z$ have been adjusted for the $n \times p$ covariate matrix $\mathbf{Z}$. For the (continuous) $\mathbf{x}$ values, we obtain $\mathbf{x}_{i.,z}$ as scaled residuals from the linear regression of $\mathbf{x}_{i.}$ on $\mathbf{Z}$. For continuous $\mathbf{y}$, we obtain $\mathbf{y}_z$ in the same manner. Thus $S_{i,z}$ is proportional to the partial correlation between $\mathbf{x}_{i.}$ and $\mathbf{y}$ after adjusting for $\mathbf{Z}$.

Applying the residualization procedure to each gene, we obtain the adjusted $\mathbf{X}_z$, which is then used (along with $\mathbf{y}_z$) to compute U_z, V_z, and D_z using the methods described earlier. For large n, this procedure may be highly accurate. However, the rank of $\mathbf{X}_z$ is now $n - p - 1$, while the permutation procedure does not "know" that the data have been reduced in rank. Thus we must adjust for the reduced rank, which is especially important for small sample sizes. Another way to view the problem is that the residualized quantities are no longer exchangeable, because the actual matching of $\mathbf{X}_z$ with $\mathbf{y}_z$ produces different variability properties for the observed global statistics than for the permuted versions. In order to preserve type I error, for U_z we compare the observed $r_{U,z}$ to the standard r density, but using a "sample size" of $n-p$. Note that this approach accords with the standard r approximation, for which standard normal regression results show that $E(r^2_{U,z}) = 1/(n-p-1)$, while $E(r^2_U) = 1/(n-1)$ under the null, either unconditionally for normally distributed $\mathbf{y}$, or exactly by permutation.

For V_z, two potential correction approaches might be considered. Note that equations (2) and (3) are still equivalent when using $\mathbf{X}_{cat,z}$ and $\mathbf{y}_z$. Based on the argument above, we might adjust V_z by the factor $(n-p-1)/(n-1)$, following the argument above, on the grounds that each constituent r_j^2 in equation (3) has effectively been inflated by that same factor. We could then proceed as usual using eigenvalues of $\mathbf{X}_{cat,z}$, for which the last $p+1$ values are now zero (assuming $m_{cat} > n$). Although this approach provides the correct mean value for V_z, the variance is incorrect, as the $\{r_j^2\}$ (on the residualized data) now have stronger correlation due to the covariate adjustment. To see why, note that the number of orthonormal PCs of $\mathbf{X}_{cat,z}$ that span the space of $\mathbf{X}_{cat}$ and are orthogonal to $\mathbf{Z}$ is $n-p-1$. Thus, by the arguments of Appendix B1, only $n-p-1$ $\{r_{j,z}^2\}$ terms should be used. Thus, instead of adjusting V_z, the correct approach is to apply the weighted beta approximation in the ordinary manner, using the eigenvalues from $\mathbf{X}_{cat,z}$, but with the smaller effective sample size. Note that this approach effectively treats the component $\{r_{j,z}^2\}$ terms as if they have a larger average than that induced under permutation. The same approach to covariate correction is also used for D_z.

For dichotomous response $\mathbf{y}$, we perform the adjustment using logistic regression residuals $\mathbf{y}_z = (y_j - \frac{e^{\hat{\beta}\mathbf{Z}}}{1+e^{\hat{\beta}\mathbf{Z}}})$, for which the $p+1$ vector β is estimated via maximum-likelihood. The results of earlier sections already showed that the weighted beta approximation fit remarkably well for dichotomous data, and after residualization the new $\mathbf{y}_z$ values are typically continuous.

2.5 Survival analysis

For censored data, we propose using the martingale residuals

$$y_j = \delta_j - \widehat{\Lambda}_0(t_j)e^{\widehat{\beta}Z},$$

(Therneau & Grambsch (1990)), where δ_j is the death (i.e., non-censored) indicator for the jth individual at the observed time t_j, $\hat{\Lambda}_0$ is the estimated cumulative hazard, and $\widehat{\beta}\mathbf{Z}$ are the covariate predictions, if applicable. Martingale residuals have also been used in a version of *globaltest* (Goeman *et al.* (2005)). Once the residualization has been performed, the *safeExpress* procedure can proceed in the manner described above.

3 Results

We assessed the performance of our analytic approximations using categories for two datasets: (i) the saliva data described earlier ($n = 18$, m=54,675 probe sets), with KEGG annotation, age as a continuous response, and an $n_0 = n_1 = 9$ dichotomous response; (ii) a head and neck squamous cell carcinoma (HNSCC) dataset ($n = 35$, E-GEOD-3292, HG-U133+2, m=54,861 probe sets), in which 8 samples were infected with human papilloma virus (HPV), 27 were uninfected, and for which a single normal continuous phenotype was simulated and used throughout. We selected KEGG categories for the saliva dataset and Gene Ontology categories for the HNSCC dataset.

Tables 1 and 2 display the type I error results for U, V and D for illustrative categories. For the saliva dichotomous response, the results are based on exhaustive permutation, and for other responses are based on 10^6 random permutations. To cover the range of useful category testing thresholds, we display ratio$_\alpha$ =(true Type I error)/α, for α ranging from 10^{-1} to 10^{-4}. The results are strikingly accurate, with ratio$_\alpha$ ranging from 0.8 to 1.39, and with most values very close to 1.0. In addition, for V we show the results from the scaled central chisquare approximation, using both the naive moments and the corrected moments based on the weighted beta. For the saliva data, the naive moments are extremely poor, which is due to the fact that the first few eigenvalues do not dominate, such that the correlation of $\{r_j^2\}$ terms is highly consequential. In fact, for Table 1 the true variance of V and the naive variance estimate differ by a factor of about 8.68, and the naive scaled central chisquared approximation produces ratio$_\alpha$ near zero. Even when using the corrected

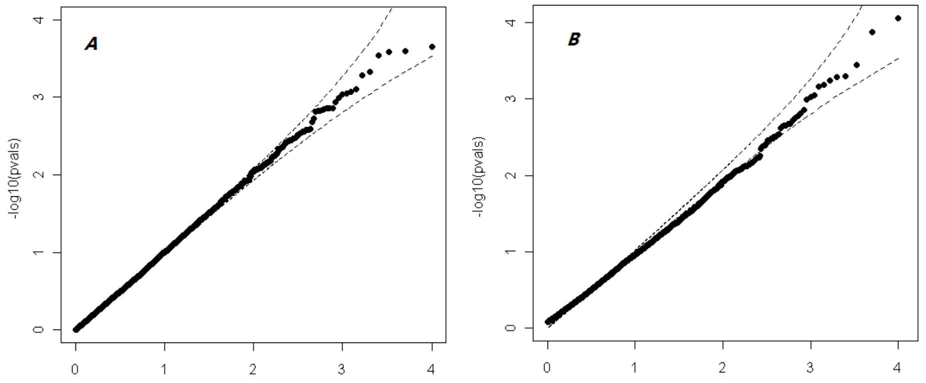


Figure 4: Observed vs. expected covariate-adjusted p-values for U_z (A) and V_z (B), KEGG:00150, salivary data, with dichotomous $\mathbf{y}$.

moments, the chisquared approximation is very poor, with ratio$_\alpha$ exceeding 2.0 or even much higher.

To demonstrate the effectiveness of the covariate adjustment, we again use KEGG:00510 (92 genes) for the saliva data. We consider the scenario which provides challenges for proper covariate adjustment: (i) a small sample size ($n = 18$); (ii) a dichotomous phenotype, and (iii) two covariates, one which is correlated with both $\mathbf{X}_{cat}$ and $\mathbf{y}$. Recall that in earlier results, the data (both expression and response) could be considered fixed, and it was sensible for each category to create a single permutation distribution to which the analytic approximations could be compared. If type I error is always controlled, conditioned on the observed data, then it follows that type I error is also controlled unconditionally. Here, however, in the presence of covariates, there is no single permutation distribution to be generated. Thus, although we still treat the gene expression data as fixed, we randomly generated covariates and the response $\mathbf{y}$ according to a model. Then we investigated the type I error behavior of the approach, which is conditional on $\mathbf{X}_{cat}$ but unconditional for $\mathbf{y}$ and $\mathbf{Z}$.

For this example, covariate z_1 was generated as $0.2\times$ $\mathbf{x}'$ of $\mathbf{X}_{cat}$, plus a $N(0,1)$ error term. Response $\mathbf{y}$ was generated using the logistic model with $logit(y) = z$, 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 simulations show good performance of the resulting p-values using the standard r approximation for $r_{U,z}$ (Figure 4A), and the weighted beta approximation for V_z (Figure 4B). Further results for V in Supplementary Figure 1 show that the naive approaches are very poor, in which the failure to adjust for covariates, or of incorrectly failing to account for the reduced rank of the weighted beta approximation, produce highly inaccurate p-values. Similar results hold for continuous phenotypes, for which an example is shown in Supplementary Figure 2.

To further illustrate the performance of our methods, we analyzed the breast cancer data of Miller et al., 2005 (GEO dataset GSE3493, Miller *et al.* (2005)), with 251 Affymetrix U133A samples and 22,215 probe sets. A total of $K = 6701$ KEGG and GO categories were examined, using both *safe* (π=10,000 permutations) and *safeExpress.* Although *safeExpress* is far faster than *safe* for individual categories, for this large number of categories *safeExpress* was only 4 times faster using Revolution R v.4.3 on a 64-bit Windows PC (16 minutes vs. 47 minutes). Separate testing using standard R indicated a 2-fold improvement using *safeExpress* (29 minutes vs. 68 minutes). However, here the primary advantage is not speed. Even with 10,000 permutations, *safe* cannot provide p-values of sufficient resolution for the K categories. If, for a particular category, none of the permutations is as extreme as the observed data, then a conservative approach is to use p-value $1/\pi$. Applying the Bonferroni procedure results in a minimum adjusted p-value $K/\pi = .67$, and the results are unlikely to be significant by any multiple test procedure. Alternately, for such a category one might report p-value $= 0$, because no permutation was extreme as that observed. However, this approach can vastly overrepresent the true significance, does not allow for proper multiple test correction, and fails to distinguish among the most significant categories.

Using *safeExpress*, we found several highly significant GO categories for two response vectors **y** under various scenarios: (i) the dichotomous p53 mutation status; (ii) p53 mutation status for estrogen-receptor positive (ER+) tumors only(n=213); (iii) p53 mutation status after adjusting for ER status in all patients, and (iv) the continuous martingale residuals for disease-free survival in all patients. The most significant categories, with Holms' adjusted p-values applied to each of U, V, and D are provided as supplementary files. Here we highlight just a few results from (iii) and (iv), with potential supporting literature. Results for (iii) included GO:0000778, Condensed Nuclear Chromosome Kinetochore (p=5.2$\times$ 10^{-15} for U) and other categories related to mitotic chromosome condensation (Chi *et al.* (2009)), and GO:0045767, Regulation of Anti-apoptosis (p=8.5 $\times$ 10^{-10} for D). The last finding accords with the role of p53 in apoptosis (Amaral *et al.* (2009)). For disease-free survival response (iv), fewer categories were significant after multiple test correction. These included GO:0051087, Chaperone Binding (p=3 $\times$ 10^{-6} for U) (Marx *et al.* (2007)), and GO:0009896 Positive Regulation of Catabolic Process (p=1 $\times$ 10^{-7} for D) (Wallace *et al.* (2000)), a category that includes *IGF-1* , widely studied for its role in breast and other cancers (Chong *et al.* (2011)) For this last result, a "safe-plot" of the ranks of local statistics S_i within the category is shown in Figure 5. Note that the genes tend to have extreme positive or negative score statistics, representing poor or good prognosis when expression is high, respectively. The ranks of the scores depart markedly from the expected uniform cdf diagonal.

The results in earlier sections had illustrated the accuracy of our approximations for individual categories. Supplementary Figure 4 compares the $-\log_{10}(p)$-values of *safe* permutation vs. the *safeExpress* approximations for U, V, and D, across all 6701 categories for the p53 mutation status response. The agreement is close, and variation for large $-\log_{10}(p)$ is largely due to sampling variation from the permutations. Note that the permutation p-values cannot be less than 10^{-4}, while the *safeExpress* values are not limited, and thus can display the true significance of each category and differentiate levels of significance among highly

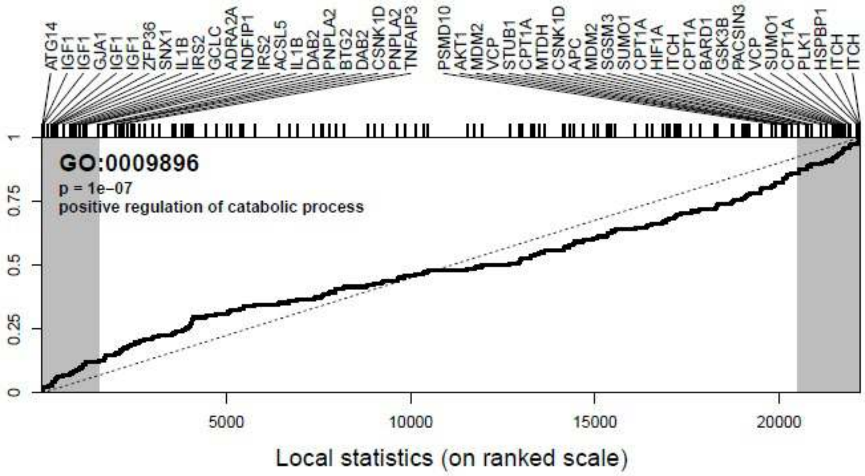


Figure 5: Safe-plot of ranks of local S statistics, using disease-free survival for the breast cancer data, GO: 0009896.

significant categories.

4 Discussion

We have demonstrated that a careful modeling approach provides highly accurate approximations to permutation p-values for gene category testing. Although confined to statistics that involve linear operations, our results include several novel techniques that are likely useful in other contexts. The kurtosis-corrected r approximation can apply generally to two-sample problems, and the weighted beta approximation applies to sums of squared score statistics, which are widely used in ensemble hypothesis testing. Finally, we are not aware that any accurate analytic approximations to mean differences of squared score sums (D, in our context) have been previously proposed.

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

Acknowledgements

Supported in part by the Gillings Statistical Genomics Innovation Lab, EPA RD83382501, NCI P01CA142538, NIEHS P30ES010126 and P42ES005948. We thank Dr. Wei Pan for pointing out the connections to the globaltest statistic.

References

Abd-Elfattah, E. F. (2009). Testing for independence: Saddlepoint approximation to associated permutation distributions. *Electronic Journal of Statistics* **3**(625-632).

Amaral, J. D., Xavier, J. M., Steer, C. J., & Rodrigues, C. M. (2009). The role of p53 in apoptosis. *Discov Med.* **9**(145-52).

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21-9**(1943-9).

Barry, W. T., Nobel, A. B., & Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics* **2(1)**(286-315).

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician* **36(3)**(153-157).

Chi, Y. H., Ward, J. M., Cheng, L. I., Yasunaga, J., & Jeang, K. T. (2009). Spindle assembly checkpoint and p53 deficiencies cooperate for tumorigenesis in mice. *Int J Cancer* **124**(1483-9).

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

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25**(631-650).

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., & Yasui, Y. (2009). Gene-set analysis and reduction. *Brief Bioinformatics* **10-1**(24-34).

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

Fisher, R. A. (1928). The General Sampling Distribution of the Multiple Correlation Coefficient. *Proc. R. Soc. A* **121**(654-673).

Fisher, R. A. (1938). The Statistical Utilization of Mutiple Measurements. *Annals of Human Genetics* **8-4**(376-386).

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., & Wright, F. A. (2010). Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets. *BMC Genomics* **11**(574).

Gatti, D. M., Shabalin, A. A., Lam, T. C., Wright, F. A., & Rusyn, I. (2009). FastMap: fast eQTL mapping in homozygous populations. *Bioinformatics* **25**(482-489).

Goeman, J. J. & Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**(9807).

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

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

Good, P. (2000). *Permutation Tests.*

Harris, P. & Peers, H. W. (1980). The local power of the efficient scores test statistic. *Biometrika* **67(3)**(525-9).

Jobson, J. D., Fienberg, S. E., & Olkin, I. (1991). *Applied Multivariate Data Analysis: Regression and Experimental Design.*

Kim, K. I. & van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* **9**(114).

Knijnenburg, T. A., Wessets, L. F. A., Reinders, M. J. T., & Shmulevich, H. (2009). Fewer permutations, more accurate P-values. *Bioinformatics* **25-12**(161-168).

Lee, S., Wright, F. A., & Zou, F. (2011). Control of Population Stratification by Correlation-Selected Principal Components. *Biometrics* **67-3**(967-974).

Liu, H., Tang, Y., & 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).

Liu, Z. J. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics* **87-1**(139-145).

Marx, C., Yau, C., Banwait, S., Zhou, Y., Scott, G. K., Hann, B., Park, J. W., & Benz, C. C. (2007). Proteasome-Regulated ERBB2 and Estrogen Receptor Pathways in Breast Cancer. *Molecular Pharmacology* **71**.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., H. P., Klaar, S., Liu, E. T., & Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**(13550-5).

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., hirschhorn, J. N., Altshuler, D., & C., G. L. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**(267-273).

Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S., & Ahlquist, P. (2007). Random set methods indentify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics* **1-1**(85-106).

Pan, W. (2009). Asymptotic Tests of Association with Multiple SNPs in Linkage Disequilibrium. *Genet. Epidemiol.* **33**(497-507).

Pollard, K. S. & van der Laan, M. J. (2005). Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data. *J. Statistical Planning and Inference* **125**(85-100).

Robinson, J. (1982). Saddlepoint approximations for the distribution of a sum of independent random variables. *J.R. Statist. Soc. B* **44**(91-101).

Schaid, D. J. & Sommer, S. S. (1994). Comparison of Statistics for Candidate-Gene Association Studies Using Cases and Parents. *Am.J. Hum. Genet.* **55**(402-409).

Srivastava, M. S. & Lee, G. C. (1984). On the Distribution of the Correlation Coefficient When Sampling from a Mixture of Two Bivariate Normal Densities: Robustness and the Effect of Outliers. *The Canadian Journal of Statistics* **12**(119-133).

Subramanian, A. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**(15545-15550).

Therneau, T. M. & Grambsch, P. M. (1990). Martingale-based residuals for survival models. *Biometrika* **77**(147-60).

Virtaneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de la Chapelle, A., & Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl Acad. Sci.* **98(3)**(1124-1129).

Wallace, H. M., Duthie, J., Duthie, J., Evans, D. M., Lamond, S., Nicoll, K. M., & Heys, S. D. (2000). Alterations in Polyamine Catabolic Enzymes in Human Breast Cancer Tissue. *Clinical Cancer Research* **6**(3657-3661).

Table 1: Performance of the p-value approximations for KEGG:00940 (18 genes), saliva dataset. Entries in the table are ratio$_\alpha$ =(true type I error table)/α.

threshold	Continuous y			Discrete y			Scaled central χ^2		Corrected SC χ^2	
α	U	V	D	U	V	D	$V_{continuous}$	$V_{discrete}$	$V_{continuous}$	$V_{discrete}$
10^{-1}	1.00	1.02	1.02	1.00	1.12	1.16	0.02	0.65	1.65	1.09
10^{-2}	1.03	1.06	1.06	1.01	1.24	1.32	0.00	0.22	4.41	1.66
10^{-3}	1.01	1.10	1.09	1.03	1.52	1.56	0.00	0.04	15.71	2.51
10^{-4}	1.09	1.17	1.11	1.03	0.82	1.65	0.00	0.00	54.71	4.11

Table 2: Performance of the p-value approximations for GO:0031012 (286 genes), HPV dataset. Entries in the table are ratio$_\alpha$ =(true type I error table)/α.

threshold	Continuous y			Discrete y			Scaled central χ^2		Corrected SC χ^2	
α	U	V	D	U	V	D	$V_{continuous}$	$V_{discrete}$	$V_{continuous}$	$V_{discrete}$
10^{-1}	1.00	1.01	1.00	1.00	0.98	1.01	0.83	0.84	0.92	0.93
10^{-2}	1.03	1.04	1.04	1.00	0.91	1.02	1.44	1.37	2.07	2.05
10^{-3}	1.05	1.06	1.02	1.00	0.82	1.03	2.88	2.41	5.71	5.19
10^{-4}	1.22	1.15	1.16	1.00	0.50	1.20	5.50	4.60	16.50	12.40

1 Appendix A1: Simple results for U and r_U

1.1 *U is monotone with r_U.*

We have

$$\begin{aligned} r_U &= corr(\mathbf{x}'_{i.}, \mathbf{y}) \\ &= \frac{\sum_j x'_j y_j - n\overline{\mathbf{xy}}}{s_{\mathbf{x}} s_{\mathbf{y}}} \\ &= \frac{U}{s_{\mathbf{x}} s_{\mathbf{y}}} \end{aligned} \tag{1}$$

because $\overline{\mathbf{x}} = 0$ and the sample standard deviations $s_{\mathbf{x}}$, $s_{\mathbf{y}}$ are constant over permutations. Therefore r_U is one-to-one with U over the permutations.

1.2 $E(r^2) = \frac{1}{n-1}$.

For any vectors $\mathbf{x}$,$\mathbf{y}$, we assume without loss of generality that they have been scaled so that

$$\sum_j x_j = 0, \sum_j x_j^2 = 1, \sum_j y_j = 0, \sum_j y_j^2 = 1.$$

For such scaling,

$$\begin{aligned} r^2 &= (\sum_j x_j y_j)^2 \\ &= \sum_j (x_j y_j)^2 + \sum_{j' \neq j} x_j y_j x_{j'} y_{j'}. \end{aligned} \tag{2}$$

For a randomly chosen j, we have $E(x_j^2) = \frac{1}{n}$, which follows from $\sum x_j^2 = 1$.

Also, for a randomly chosen pair j, $j' \neq j$,

$$
\begin{aligned}
E(x_j x_{j\prime}) &= \sum_j \sum_{j'j} x_j x_{j'} x_j x_{j'} p(x_j, x_{j'}) \\
&= \sum_j x_j(-x_j)\frac{1}{n(n-1)} \\
&= \frac{1}{n(n-1)} \sum_j -x_j^2 \\
&= -\frac{1}{n(n-1)}. \qquad (3)
\end{aligned}
$$

For the permutations, we consider the $(n!)^2$ arrangements of fixed $\mathbf{x}$ and $\mathbf{y}$, and let E_Π denote expectation over the permutations. We have

$$
\begin{aligned}
E_\Pi(r^2) &= E_\Pi[(\sum_j x_j y_j)^2] \\
&= E_\Pi(\sum_j (x_j y_j)^2) + E_\Pi(\sum_j \sum_{j' \neq j} x_j y_j x_{j'} y_{j'}) \\
&= n\frac{1}{n^2} + n(n-1)\big(-\frac{1}{n(n-1)}\big)^2 \\
&= \frac{1}{n-1}. \qquad (4)
\end{aligned}
$$

1.3 The distribution of r when one of the two variables is normal

The following result is widely known, but a number of sources give the incorrect impression that it requires *bivariate* normality. As described in Lehmann and Romano (Problem 564, page 207), for fixed x_1,..., x_n, and normal random $\mathbf{y}$ and true correlation $= 0$, the distribution of $(n-2)r/\sqrt{1-r^2}$ follows a central t with $n-2$ degrees of freedom. As $\mathbf{x}$ is arbitrary, it follows that the same result holds unconditionally for random $\mathbf{x}$. Furthermore, $corr(x, y) = corr(y, x)$, and so the role of x and y is arbitrary. It follows that only one of the two variables need be normal in order for r to follow the standard density.

In our setting, we are sampling without replacement, and thus the values are slightly dependent. Nevertheless, the approximation appears to be accurate as long as either $\mathbf{x}$ or $\mathbf{y}$ is nearly (empirically) normal.

2 Appendix A2: The saddlepoint approximation for r_U when y is dichotomous

The problem of permutation-based comparison of two samples was considered by Robinson (1982), which correspond in our case to dichotomous $\mathbf{y}$ (assumed to consist of n_0 zeros and n_1 ones). Here we follow the Robinson notation, with modifications to accord with our notation. New standardized quantities are computed,

$$a_j = (x'_j - \overline{\mathbf{x}}')/[\sum_{i=1}^{n}(x'_j - \overline{\mathbf{x}}')^2]^{\frac{1}{2}},$$

and the Robinson test statistic is

$$t_n = w_n^{-\frac{1}{2}} \sum_{j \in \{y=1\}} a_j,$$

where $w_n = npq/(n-1)$ and $p = n_0/n$, $q = 1-p$. As $\sum_{j \in \{y=1\}} x'_j = \sum_{j=1}^{n} x'_j y_j$, it is simple to show (as in Appendix A1) that t_n and r_U are equivalent statistics.

The saddlepoint approximation involves computing quantities

$$m_n = w_n^{-\frac{1}{2}} \sum_j a_j K'_j$$

and

$$\sigma_n^2 = w_n^{-1}\{\sum_j a_j^2 K'' - (\sum_j a_j K'')^2 / \sum_j K''_j\},$$

where (i) $K(x) = log(pe^{qx} + qe^{-px})$; (ii) K_j, K'_j and K''_j denote $K(x), K'(x)$ and $K''(x)$ evaluated at $x = ua_j w_N^{-1/2} + \alpha(u)$, where $\alpha(u)$ solves $\sum_j K'_j = 0$ and u is described below. Upon differentiation we have $K' = \frac{1}{pe^{qx}+qe^{-px}}(pqe^{qx} - pqe^{-px})$ and $K'' = \frac{(pe^{qx}+qe^{-px})(pq^2e^{qx}+p^2qe^{-px})}{(pe^{qx}+qe^{-px})^2} - \frac{(pqe^{qx}-pqe^{-px})^2}{(pe^{qx}+qe^{-px})^2}$.

The choice of u simultaneously solves $m_n(u) = t_n$ (where t_n is the realized statistic of the permutation-random T_n) and $\sum_j K'_j = 0$, and u is obtained numerically using *optim*() in R. Finally, the saddlepoint approximate tail probability is given by

$$Q_n(u)\text{exp}(-um_n + (1/2)u^2\sigma_n^2)(1 - \Phi(u\sigma_n)),$$

where Q is the moment generating function for T_n, σ^2 is computed for u, and Φ is the cdf of a standard normal density. Knowledge of the moment generating function technically requires knowing the desired distribution of T_n. Thus, in typical saddlepoint applications, the moment generating function must be further approximated in order that the approach be practically useful. However, as we have performed exhaustive permutation in our examples, we can exactly compute the mgf

$$Q_n(u) = \frac{1}{|\Pi|}\sum_{\Pi}\text{exp}(ut_{n,\Pi}),$$

which we use in our examples order to present a best-case scenario for the saddlepoint.

We compared our results to the two two-sample datasets in Robinson (1982), and obtained identical results as reported by the author.

3 Appendix A3: Kurtosis of r vs. kurtosis of $\mathbf{x}$

The (excess) kurtosis is invariant under linear transformation, hence

$$kurtosis_{\Pi}(U) = kurtosis_{\Pi}(r_U),$$

and we emphasize that Π signifies a population that is created conditional on the data $\mathbf{x}'$ and $\mathbf{y}$. The main result of this section is that, when $\mathbf{y}$ is dichotomous, the kurtosis of r_U over permutations is determined entirely by the kurtosis of the observed n-vector $\mathbf{x}'$. In order to avoid confusion over "population" and "sample" quantities, here the kurtosis will always refer to a population parameter. For this purpose the observed $\mathbf{x}'$ will be considered a population of size n, and probability $1/n$ assigned to each element.

As $U \propto \sum_{j=1}^{n} x'_j y_j = \sum_{j\in\{y=1\}} x'_j$, it follows that U is the same as the summation of a random sample of size n_1 from a population of size n. Moreover $U = n_1\frac{\sum_{j\in\{y=1\}} x'_j}{n_1} = n_1\widehat{\mu}$, where $\widehat{\mu}$ is the sample mean of drawing n_1

values from the population of size n, without replacement. Thus the kurtosis of $\hat{\mu}$ is equal to that of r_U. Formulas for moments of $\hat{\mu}$ up to order 4 are provided in Kendall & Stuart (1969) Chapter 12, using cumulant notation and "symmetric k" statistics to simplify computation. We denote the kth central moment of $\mathbf{x}'$ as

$$\mu_k = (1/n)\sum_j (x'_j - \overline{\mathbf{x}}')^k.$$

Also, using the following definition (Kendall & Stuart (1969), chapter 12, formula 12.29),

$$K_4 = \frac{n_1^2}{(n_1-1)(n_1-2)(n_1-3)}\{(n_1+1)\mu_4 - 3(n_1-1)\mu_2^2\},$$

we have

$$\begin{aligned} E((\hat{\mu}-\mu)^4) &= K_4\{(\frac{1}{n_1^3}-\frac{1}{n^3}) - \frac{4}{n}(\frac{1}{n_1^2}-\frac{1}{n^2}) + \frac{6}{n^2}(\frac{1}{n_1}-\frac{1}{n})\} \\ &+ 3\frac{n-1}{n+1}(\sigma^4 - \frac{K_4}{n})\{(\frac{1}{n_1^2}-\frac{1}{n^2}) - \frac{2}{n_1}(\frac{1}{n_1}-\frac{1}{n})\}, \end{aligned} \tag{5}$$

where we follow the Kendall notation $\sigma^2 = \mu_2 n/(n-1)$, and a simple finite-sampling result shows that $\text{var}(\hat{\mu}) = \sigma^2(\frac{1}{n_1} - \frac{1}{n})$. Finally, we compute the kurtosis of the sample mean (and therefore for r_U) as

$$\frac{E\big((\hat{\mu}-\mu)^4\big)}{\text{var}^2(\hat{\mu})} - 3.$$

4 Appendix A4: An improved approximation to r for dichotomous y

We will denote the standard r density as $f_n(r)$, and the subscript signifies the dependence on n. The variance of r is always $1/(n-1)$ under permutation, and if $n_0 \approx n_1$, then the permutation distribution will be nearly symmetric (exactly so if $n_0 = n_1$). Thus we focus on the kurtosis of r, which is solved in Appendix A3 as a function of kurtosis($\mathbf{x}'$), to provide the basis for an improved fit. To improve upon the standard approximation, we consider the impact of a hypothetical outlier x_j in the comparison of a continuous x vector to dichotomous y. Suppose that in the absence of the jth observation, the standard r approximation was accurate, so that $r \sim f_{n-1}(r)$. Then in the presence of the outlier, we model the density as being "offset" by a value c, according to whether the outlier coincides with $y = 0$ or $y = 1$, with equal probability symmetrically about zero. Thus the new density would take the form $f_{new}(r) = \frac{1}{2}\{f_{n-1}(r - c) + f_{n-1}(r + c)\}$. However, the variance of r should still be constrained to $1/(n-1)$, and we replace f_{n-1} with f_n in order that the standard density hold in the special case that $c = 0$. The final approximation is

$$f_{new}(r) = \frac{K}{2}\{f_n(rK - c) + f_n(rK + c)\}$$

where $K = \sqrt{1 + c^2(n-1)}$. The new approximation accords with f_n when $c = 0$, and tends to have lower (i.e. more negative) kurtosis as c increases, although the kurtosis is not strictly monotone with c near zero. Thus if there are multiple real solutions for a desired kurtosis, we chose the c solution nearest zero.

Below we solve for c for a given kurtosis(r). We note that the motivation above envisions a new random variable $r_{new} = r + cB$, where r follows the standard density, and B is an independent random variable assuming the values -1 and 1 each with probability 1/2. Note that in this setup r_{new} has not been rescaled to have the correct variance, but this is irrelevant, as the kurtosis is invariant to linear scaling. For this subsection we will let k denote the desired kurtosis(r), and we will solve for c such that k=kurtosis(r_{new}). The expectations are over exhaustive permutations. We have $k = \text{kurtosis}(r_{new}) = \frac{E((r_{new}-E(r_{new}))^4)}{\sigma^4} - 3 = \frac{E(r_{new}^4)}{(var(r)+c^2)^2} - 3$

$$\begin{aligned}
E(r_{new}^4) &= E(r^4 + 4r^3cB + 6r^2c^2B^2 + 4rc^3B^3 + c^4B^4) && (6)\\
&= E(r^4 + 6r^2c^2B^2 + c^4B^4) && (7)\\
&= E(\frac{3}{n^2-1} + 6c^2B^2\frac{1}{n-1} + c^4B^4) && (8)\\
&= \frac{3}{n^2-1} + 6c^2\frac{1}{n-1} + c^4 && (9)
\end{aligned}$$

Therefore

$$k = \frac{\frac{3}{n^2-1} + \frac{6}{n-1}c^2 + c^4}{(\frac{1}{n-1} + c^2)^2} - 3 \tag{10}$$

$$= \frac{3}{n^2-1} + \frac{6}{n-1}c^2 + c^4 \quad = \quad (k+3)[\frac{1}{(n-1)^2} + c^4 + 2\frac{c^2}{n-1}] \tag{11}$$

and

$$(k+2)c^4 + \frac{2k}{n-1}c^2 + \{\frac{k+3}{(n-1)^2} - \frac{3}{n^2-1}\} = 0.$$

The quadratic solution for c^2 is

$$c^2 \quad = \quad \frac{-\frac{k}{n-1} \pm \sqrt{(k/(n-1))^2 - (k+2)(k+3)/(n-1)^2 + 3(k+2)/(n^2-1)}}{k+2} \tag{12}$$

$$= \quad \frac{-\frac{k}{n-1} \pm \sqrt{\frac{-5k-6}{(n-1)^2} + 3\frac{k+2}{n^2-1}}}{k+2}. \tag{13}$$

We solve for $c = \sqrt{c^2}$, and use the real root nearest zero, or $c = 0$ if there are no real roots.

5 Appendix B1: The weighted beta distribution

Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ be i.i.d normally distributed with variance σ^2 and independent of all the orthonormal basis $\mathbf{p_{.j}}$, $j = 1, ..., n$ in R^n. Denote $(y_1 + y_2 + ... + y_n)/n$ as $\overline{\mathbf{y}}$, and let $b_k = \sum_j y_j \mathbf{p_{1j}}$. Then since the mean of $\mathbf{p_{.k}}$ is zero, the sample correlation between $\mathbf{y}$ and $\mathbf{p_{.k}}$ for $k \geq 1$ is given by

$$r_k \equiv b_k / \sqrt{S_n^2},$$

where $S_n^2 = \sum_i (y_i - \overline{y})^2$.

Since $\mathbf{y} = \sum_{k=0}^{n-1} < \mathbf{y}, \mathbf{p_{.k}} > \mathbf{p_{.k}}$, where $< \mathbf{y}, \mathbf{p}_{.k} >$ is the inner product of $\mathbf{y}$ and $\mathbf{p}_{.k}$, it is easy to see

$$\mathbf{y} - \overline{\mathbf{y}} = \sum_{k=1}^{n-1} b_k \mathbf{p}_{.k}.$$

Moreover,

$$S_n^2/\sigma^2 = ||\mathbf{y} - \overline{\mathbf{y}}||^2/\sigma^2 = \sum_{k=1}^{n-1} b_k^2/\sigma^2.$$

On the other hand, conditioning on $\mathbf{p_{.1}}$, ..., $\mathbf{p_{.n}}$, each b_k follows a normal distribution with mean zero, variance σ^2 and $\mathrm{Cov}(b_k, b_l) = 0$. This implies that each $Z_k \equiv b_k^2/\sigma^2, k = 1, ..., n-1$ are i.i.d and distributed χ_1^2.

Therefore, using a fact Gupta & Nadarajah (2004)

$$\frac{\chi_m^2}{\chi_m^2 + \chi_l^2} \sim Beta(m/2, l/2),$$

we have

$$r_k^2 = b_k^2/(S_n^2) = \frac{Z_k}{\sum_{j=1}^{n-1} Z_j} \sim Beta(1/2, (n-2)/2),$$

then for any subset $\Omega \subset \{1, ..., n-1\}$ which doesn't contain k,

$$\frac{r_k^2}{1 - \sum_{j \in \Omega} r_j^2} = \frac{Z_k}{\sum_{j \notin \Omega} Z_j} \sim Beta(1/2, (n - |\Omega| - 2)/2).$$

Since the distributions are independent of the $\mathbf{p}_{.k}$'s, these distributions are also unconditional distributions.

The joint density is

$$p(r_1^2, r_2^2, ..., r_{n-1}^2) = p(r_1^2)p(r_2^2|r_1^2)...p(r_{n-1}^2|r_1^2, r_2^2, ..., r_{n-2}^2).$$

Following the results above, the product includes terms

$$p(r_j^2|r_1^2, ..., r_{j-1}^2) = \frac{1}{d_j \mathrm{B}(1/2, (1/2)(n-j-1))} (r_j^2/d_j)^{-1/2} (1 - r_j^2/d_j)^{(1/2)(n-j-3)},$$

where $d_j = 1 - \sum_{j'=1}^{j-1} r_{j'}^2$ and B() denotes the beta function.

Although sums of subsets of these correlated $\{r_j^2\}$ are themselves distributed as beta, we are not aware that the density of the weighted sum $V = n \sum_{j=1}^{n} \lambda_j r_j^2$ has a tractable form. Numerical integration over all $n-2$ free r_j^2 terms is infeasible, even for modest n. In practice, a shifted gamma density, with shape, scale, and location parameters, provides a reasonable approximation to V. However, the quality of the fit tends to degrade in the extreme right tail. As a compromise between these extremes, we have chosen to fit the joint beta density of $\{r_1^2, r_2^2\}$ over a numeric grid, while fitting $V - \lambda_1 r_1^2 - \lambda_2 r_2^2$ using a shifted gamma density. The first three moments for the gamma fit are obtained using the weighted beta moments as described in Appendix B3 below, but using only eigenvalues $\lambda_3, ..., \lambda_{n-1}$.

The final p-value is obtained as

$$P(V \geq v_{observed}) = \int_0^1 \int_0^1 p(r_1^2, r_2^2) P(V \geq v_{observed} | r_1^2, r_2^2) dr_1^2 \; dr_2^2$$

and using the shifted gamma as the integrand. This simple grid approach is most accurate in the right tail, where accuracy is most desired, and for typical sample sizes can be computed for over 1000 gene categories in a few minutes.

6 Appendix B2: Correlation of squared correlations

To obtain moments for the approximation of V, and to better understand the behavior of the global statistic D, here we consider the correlation of paired quantities $r_{P_1}^2 = corr^2(P_1, \mathbf{y})$ v.s. $r_Q^2 = corr^2(Q, \mathbf{y})$, expressed in terms of the observed correlations of vectors P_1 and Q with each other.

The derivations assume $\mathbf{y}$ is random and normally distributed, for fixed P_1 and Q.

Suppose P_1 and Q are normalized vectors, not necessarily orthogonal. We claim that

$$Corr^2(r_{P_1}^2, r_Q^2) = -\frac{1}{n-2} + \frac{n-1}{n-2} corr^2(P_1, Q),$$

where $Corr^2(r_{P_1}^2, r_Q^2)$ is determined over random $\mathbf{y}$, while $corr^2(P_1, Q)$ is the sample Pearson correlation between P_1 and Q. Without loss of generality, we can choose another vector, orthogonal to P_1 (which we'll denote P_2), such that $Q = aP_1 + bP_2$.

We have

$$\begin{aligned} r_Q^2 &= corr^2(Y, a\ P_1 + b\ P_2) \\ &= (a\ corr(Y, P_1) + b\ corr(Y, P_2))^2 \\ &= a^2 corr^2(Y, P_1) + b^2 corr^2(Y, P_2) + 2abScorr(Y, P_1) corr(Y, P_2) \\ &= a^2 r_{P_1}^2 + b^2 r_{P_2}^2 + 2ab\ r_{P_1} r_{P_2}. \end{aligned} \tag{14}$$

Using previous results on Beta densities for random r^2, we have

$$\begin{aligned} E(r_{P_1}^3 r_{P_2}) - E(r_{P_1} r_{P_2})E(r_{P_1}^2) &= E(E(r_{P_1}^3 sign(r_{P_2})(B_1^{\frac{1}{2}}(1-r_{P_1}^2)^{\frac{1}{2}})|P_1)) \\ &- E(r_{P_1} sign(r_{P_2})B_1^{\frac{1}{2}}(1-r_{P_1}^2)^{\frac{1}{2}})E(r_{P_1}^2) \\ &= E(r_{P_1}^3\{(1-r_{P_1}^2)^{\frac{1}{2}})\}sign(r_{P_2})E(B_1^{\frac{1}{2}}) \\ &- E(r_{P_1}^3\{(1-r_{P_1}^2)^{\frac{1}{2}})\}sign(r_{P_2})E(B_1^{\frac{1}{2}}) \\ &= 0. \end{aligned} \tag{15}$$

Thus

$$\begin{aligned} Corr^2(r_P^2, r_Q^2) &= a^2 + b^2(-\frac{1}{n-2}) + 2abcorr(r_{P_1}r_{P_2}, r_{P_2}^2) \\ &= a^2 + (1-a^2)(-\frac{1}{n-2}) + 2ab\{E(r_{P_1}^3 r_{P_2}) - E(r_{P_1}r_{P_2})E(r_{P_2}^2)\} \\ &= -\frac{1}{n-2} + \frac{n-1}{n-2}a^2 \\ &= -\frac{1}{n-2} + \frac{n-1}{n-2}corr^2(P_1, Q). \end{aligned} \tag{16}$$

Note that if Q is orthogonal to P_1, then the correlation of squared correlations is $-1/(n-2)$, which proves the desired result for any r_j^2, $r_{j'}^2$ in Equation (3). In contrast to the result of Appendix A1 1.2, which is exact over permutations, the result of this section is not exact, as it relies on the correlated beta distribution. However, simulations as described in the text indirectly show that the approximation has high accuracy, as the distributional approximation for V relies crucially on this correlation.

7 Appendix B3: The first three moments of the weighted beta approximation to V

The expectations computed in this section are unconditional, for which both $\mathbf{y}$ and the PC matrix $\mathbf{P}$ are considered to be random (and multivariate normal). However, the eigenvalues are considered fixed.

Since $E((r_1^2)^2) = \text{var}(r_1^2) + E^2(r_1^2)$, and a standard beta distributional result is that $E((r_1^2)^k) = \frac{\alpha+k-1}{\alpha+\beta+k}E((r_1^2)^{k-1})$ where r_1^2 follows a beta density with $\alpha = 1/2$, $\beta = (n-2)/1$, we have the moments

$$\begin{aligned} E(r_1^2) &= \frac{1/2}{1/2+(n-2)/2} \\ &= \frac{1}{n-1} \end{aligned} \tag{17}$$

$$\begin{aligned} E((r_1^2)^2) &= \frac{1/2+2-1}{1/2+(n-2)/2+2-1} E(r_1^2) \\ &= \frac{3}{n^2-1}. \end{aligned} \tag{18}$$

$$\begin{aligned} E((r_1^2)^3) &= \frac{1/2+3-1}{1/2+(n-2)/2+3-1} E((r_1^2)^2) \\ &= \frac{5}{n+3}\frac{3}{n^2-1} \end{aligned} \tag{19}$$

$$\begin{aligned} E(r_1^2 - 2(r_1^2)^2 + (r_1^2)^3) &= E(r_1^2) - 2E((r_1^2)^2) + E((r_1^2)^3) \\ &= \frac{1}{n-1} - 2\frac{3}{n^2-1} + \frac{5}{n+3}\frac{3}{n^2-1} \\ &= \frac{n(n-2)}{(n+3)(n^2-1)} \end{aligned} \tag{20}$$

Also, for B_1 as defined in Section 2 of the manuscript,

$$\begin{aligned} E(B_1) &= \frac{1/2}{1/2+(n-3)/2} \\ &= \frac{1}{n-2} \end{aligned} \tag{21}$$

$$\begin{aligned} E(B_1^2) &= E(B_1)\frac{1/2+1}{1/2+(n-3)/2+1} \\ &= \frac{3}{n}\frac{1}{n-2} \\ &= \frac{n-3}{n(n-2)} \end{aligned} \tag{22}$$

$$\begin{aligned} E(B_1 - B_1^2) &= E(B_1) - E(B_1^2) \\ &= \frac{1}{n-2} - \frac{3}{n}\frac{1}{n-2} \end{aligned} \tag{23}$$

$$\begin{aligned} E(B_2) &= \frac{1/2}{1/2 + (n-4)/2} \\ &= \frac{1}{n-3} \end{aligned} \tag{24}$$

We also have

$$\begin{aligned} E(r_1^2 r_2^2) &= E(r_1^2 B_1 (1 - r_1^2) \\ &= E(E(r_1^2 B_1 (1 - r_1^2) | r_1^2) \\ &= E(E(B_1 (r_1^2 - r_1^4))) \\ &= E(E(B_1) E(r_1^2 - r_1^4)) \\ &= E(\frac{1}{n-2}(\frac{1}{n-1} - \frac{3}{n^2-1})) \\ &= \frac{1}{n^2-1}, \end{aligned} \tag{25}$$

and

$$\begin{aligned} E(r_1^2 r_2^2 r_3^2) &= E(r_1^2 B_1 (1 - r_1^2) B_2 (1 - r_1^2 - r_2^2)) \\ &= E(r_1^2 (1 - r_1^2) B_1 B_2 (1 - r_1^2 - B_1 (1 - r_1^2))) \\ &= E(r_1^2 (1 - r_1^2)^2 B_1 (1 - B_1) B_2) \\ &= E(r_1^2 - 2(r_1^2)^2 + (r_1^2)^3) E(B_1 - B_1^2) E(B_2) \\ &= \frac{n(n-2)}{(n+3)(n^2-1)} \frac{n-3}{n(n-2)} \frac{1}{n-3} \\ &= \frac{1}{(n+3)(n^2-1)}. \end{aligned} \tag{26}$$

The correlation between different r_j^2 is

$$\begin{aligned} corr(r_j^2, r_{j'}^2) &= \frac{E(r_j^2 r_{\mathrm{j}'}^{\lambda} 2_{j'}) - E(r_j^2)E(r_{j'}^2)}{s_{r_j^2} s_{r_{j'}^2}} \\ &= \frac{\frac{1}{n^2-1} - \frac{1}{(n-1)^2}}{\frac{2(n-2)}{(n-1)^2(n+1)}} \\ &= -\frac{1}{n-2}. \end{aligned}$$

The first moment of V is

$$\begin{aligned} E(n\sum_{j=1}^{n} r_j^2 \lambda_j) &= n\sum_{j=1}^{n} \lambda_j E(r_j^2) \\ &= \frac{n}{n-1}\sum_{j=1}^{n} \lambda_j \\ &= \frac{m_{cat} n}{n-1}. \end{aligned} \tag{27}$$

The second moment of V is

$$\begin{aligned} E\big((n\sum_{j=1}^{n} r_j^2 \lambda_j)^2\big) &= n^2 E(\sum_{j=1}^{n} (\lambda_j)^2 (r_j^2)^2) + 2n^2 E(\sum_{j=1}^{n}\sum_{j\prime=1}^{n} r_j^2 \lambda_j r_j^2\prime \lambda_{\mathrm{j}'}) \\ &= n^2 \sum_{j=1}^{n} \lambda_j^2 E((r_j^2)^2) + 2n^2 \sum_{j=1}^{n}\sum_{j\prime=1}^{n} \lambda_j \lambda_{\mathrm{j}'} E(r_j^2 r_j^2\prime) \\ &= \sum_{j=1}^{n} \lambda_j^2 \frac{3n^2}{n^2-1} + \sum_{j=1}^{n}\sum_{j\prime=1}^{n} \lambda_j \lambda_{j\prime} \frac{2n^2}{n^2-1} \end{aligned} \tag{28}$$

Also,

$$\begin{aligned} E((r_1^2)^2 r_2^2) &= E(E(r_1^2)^2 r_2^2 | r_1^2) \\ &= E(E((r_1^2)^2 B_1 (1 - r_1^2) | r_1^2)) \\ &= E(B_1) E(r_1^4 - r_1^6) \\ &= \frac{1}{n-2}(\frac{3}{n^2-1} - \frac{5}{n+3}\frac{3}{n^2-1}) \\ &= \frac{1}{n-2}\frac{3}{n^2-1}\frac{n-2}{n+3}. \end{aligned} \tag{29}$$

$$
\begin{aligned}
E(\sum_{j=1}^{n} r_j^2\lambda_j)^3 &= E(\sum_{j=1}^{n}(r_j^2)^3\lambda_j^3) \\
&+ 3E(\sum_{j=1}^{n}\sum_{j\neq j\prime=1}^{n}(r_j^2)^2 r_{j\prime}^2\lambda_j^2\lambda_{j\prime}) \\
&+ E(\sum_{j=1}^{n}\sum_{j\prime=1}^{n}\sum_{j\prime\prime=1}^{n} r_j^2 r_{j\prime}^2 r_{j\prime\prime}^2\lambda_j\lambda_{j\prime}\lambda_{j\prime\prime}) \\
&= \sum_{j=1}^{n}\lambda_j^3 E((r_j^2)^3) + 3\sum_{j}\sum_{j\prime}\lambda_j^2\lambda_{\jmath\prime}E((r_j^2)^2 r_j^2\prime) \\
&+ \sum_{j=1}^{n}\sum_{j\prime=1}^{n}\sum_{j\prime\prime=1}^{n}\lambda_j\lambda_{j\prime}\lambda_{j\prime\prime}E(r_j^2 r_{j\prime}^2 r_{j\prime\prime}^2) \\
&= \frac{5}{n+3}\frac{3}{n^2-1}\sum_{j=1}^{n}\lambda_j^3 + 3\frac{1}{n-2}\frac{3}{n^2-1}\frac{n-2}{n+3}\sum_{j}\sum_{j\prime}\lambda_j^2\lambda_{\jmath\prime} \\
&+ \sum_{j=1}^{n}\sum_{j\prime=1}^{n}\sum_{j\prime\prime=1}^{n}\lambda_j\lambda_{j\prime}\lambda_{j\prime\prime}\frac{1}{(n+3)(n^2-1)} \qquad (30)
\end{aligned}
$$

Hence the non-central third moment of V is $E\big((n\sum_{j=1}^{n} r_j^2\lambda_j)^3\big)$, which is

$$
n^3(\frac{5}{n+3}\frac{3}{n^2-1}\sum_{j=1}^{n}\lambda_j^3+\frac{9}{(n^2-1)(n+3)}\sum_{j}\sum_{j\prime}\lambda_j^2\lambda_{\jmath\prime}+\sum_{j=1}^{n}\sum_{j\prime=1}^{n}\sum_{j\prime\prime=1}^{n}\lambda_j\lambda_{j\prime}\lambda_{j\prime\prime}\frac{1}{(n+3)(n^2-1)}).
$$

Claim :

$$
\sum_{j=1}^{n}\lambda_j = m_{cat}.
$$

Proof: Since $\mathbf{X}$ is normalized, $\sum_{j=1}^{n} x_{ij}^2 = 1$ and $\sum_{j=1}^{n} x_{ij} = 0$. We have that $\mathbf{X}_{cat}\mathbf{X}_{cat}^T$ is a matrix with diagonal entries 1, so $trace(\mathbf{X}_{cat}\mathbf{X}_{cat}^T) = m_{cat}$. A standard result holds that the singular value decomposition of X_{cat} yields singular values that are the square roots of the eigenvalues of $\mathbf{X}_{cat}\mathbf{X}_{cat}^T$, implying $trace(\mathbf{X}_{cat}\mathbf{X}_{cat}^T) = \sum_{j=1}^{n}\lambda_j = m_{cat} = m_{cat}$.

8 Appendix C: Extension of the weighted beta approach to D

Suppose $\mathbf{A}$ is the weighted gene expression matrix, i.e. $\mathbf{A} = \mathbf{WX}$, where

$$\mathbf{W} = diag[\underbrace{\frac{1}{\sqrt{m_{cat}}}, \frac{1}{\sqrt{m_{cat}}}, ..., \frac{1}{\sqrt{m_{cat}}}}_{m_{cat}}, \underbrace{\frac{1}{\sqrt{m_{comp}}}\mathcal{I}, \frac{1}{\sqrt{m_{comp}}}\mathcal{I}, ..., \frac{1}{\sqrt{m_{comp}}}\mathcal{I}}_{m_{comp}}].$$

Here we use $\mathcal{I}$ to represent $\sqrt{-1}$, to avoid confusion with the gene index i. It is simple to show that $D = \frac{Vcat}{m_{cat}} - \frac{Vcomp}{m_{comp}} = \sum_i (w_i S_i)^2$, and thus we have the desired form, similar to Equation (2).

According to the spectral decomposition, the symmetric matrix $\mathbf{A}^T\mathbf{A}$ can be decomposed as $\Phi\Psi\Phi^T$, where Φ is an orthogonal matrix, and here assumed orthonormal without loss of generality. Since the vectors $\phi_{.j}$, j=1,2,...,n form an orthonormal basis in R^n space, and $r_j = cor(\phi_{.j}, y) = \frac{\phi_{.j}^T y}{SD_{\phi_{.j}} SD_y}$, it can be shown that $y = \phi_{.j} r_j s_{\phi_{.j}} s_y$.

$$\begin{aligned}
D &= \frac{\sum_{i=1}^{m} (A_{i.}^T y)^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= \frac{\sum_{i=1}^{m} (y^T A_{i.} A_{i.}^T y)}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= \frac{\sum_{i=1}^{m} y^T (\sum_{j=1}^{n} r_j \phi_{ij} \phi_{ij}^T) y}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= \frac{\sum_{i=1}^{m} (\sum_{j=1}^{n} \psi_j \phi_{ij} \phi_{ij}^T) y^T y}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= \frac{\sum_{j=1}^{n} var(\phi_{.j}) var(y) (\sum_{i=1}^{m} \psi_j \phi_{ij} \phi_{ij}^T) (\phi_{.j} r_j)^T (\phi_{.j} r_j)}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= \frac{\sum_{j=1}^{n} \frac{1}{n-1} var(y) \psi_j r_j^T \phi_{.j}^T \phi_{.j} r_j}{\sum_{j=1}^{n} (y_j - \overline{y})^2 / n} \\
&= n \sum_{j=1}^{n} \psi_j r_j^T \phi_{.j}^T \phi_{.j} r_j \\
&= n \sum_{j=1}^{n} \psi_j r_j^2
\end{aligned}$$

Claim:

$$\sum_{j=1}^{n} \gamma_j = 0$$

We have constructed a new matrix $\mathbf{A}= \begin{bmatrix} \frac{\mathbf{X}_{cat}}{\sqrt{m_{cat}}} \\ \frac{\mathbf{X}_{out}}{\sqrt{out}} i \end{bmatrix}$,

$$\begin{aligned}
\mathbf{A}^T \mathbf{A} &= \begin{bmatrix} \frac{\mathbf{X}_{cat}}{\sqrt{m_{cat}}} & \frac{\mathbf{X}_{out}}{\sqrt{comp}} i \end{bmatrix} \begin{bmatrix} \frac{\mathbf{X}_{cat}}{\sqrt{m_{cat}}} \\ \frac{\mathbf{X}_{comp}}{\sqrt{comp}} i \end{bmatrix} && (31) \\
&= \begin{bmatrix} \frac{\mathbf{X}_{cat}^T X_{cat}}{m_{cat}} & 0 \\ 0 & -\frac{\mathbf{X}_{comp}^T \mathbf{X}_{comp}}{m_{comp}} \end{bmatrix} && (32) \\
& && (33)
\end{aligned}$$

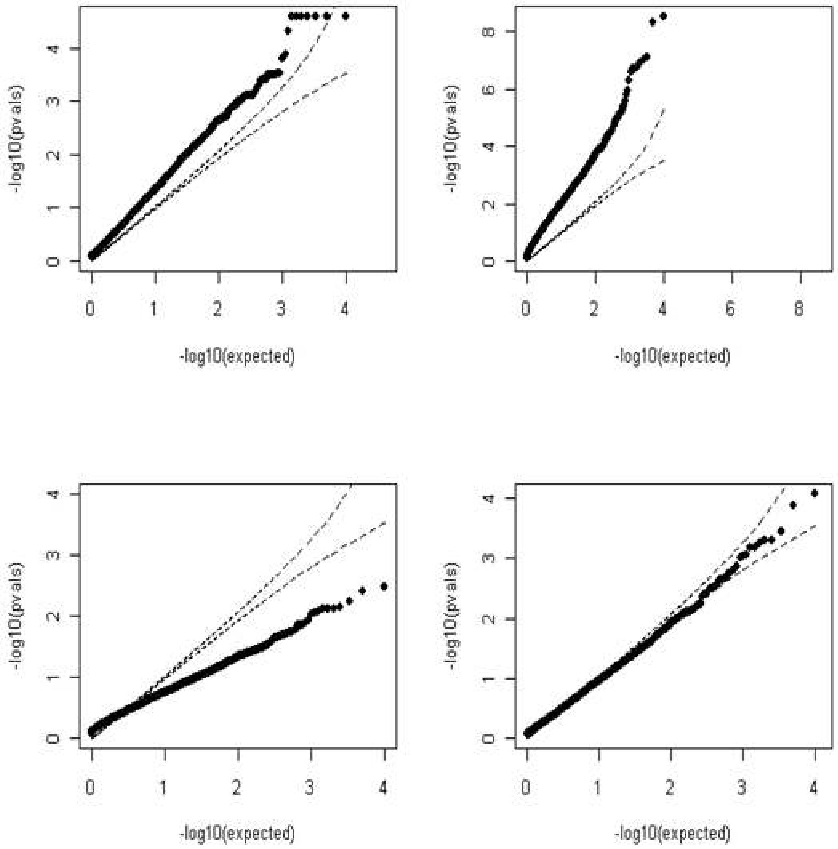
We have

$$\begin{aligned} trace(\mathbf{A}^T\mathbf{A}) &= \sum(diag(\mathbf{A}^T\mathbf{A})) \\ &= \sum(diag(\mathbf{X}_{cat}^T\mathbf{X}_{cat}))/m_{cat} - \sum(diag(\mathbf{X}_{comp}^T\mathbf{X}_{comp}))/m_{comp} \\ &= 1-1 \\ &= 0. \end{aligned} \tag{34}$$

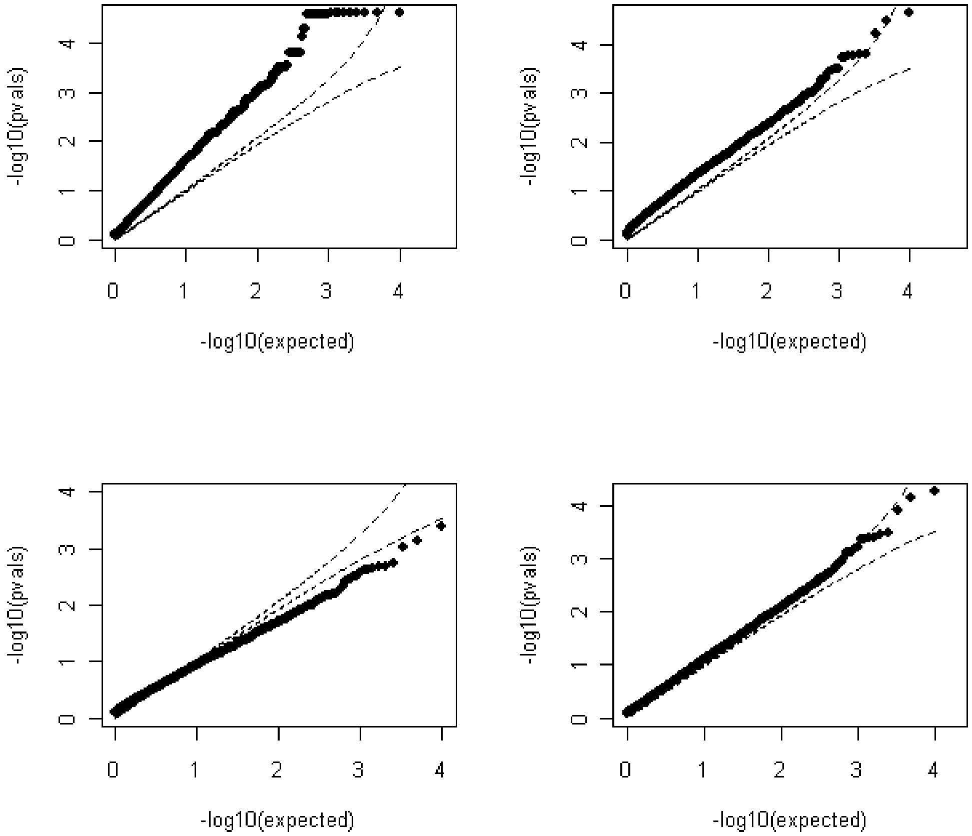
References

Gupta, A. K. & Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications, 1nd Edition.*

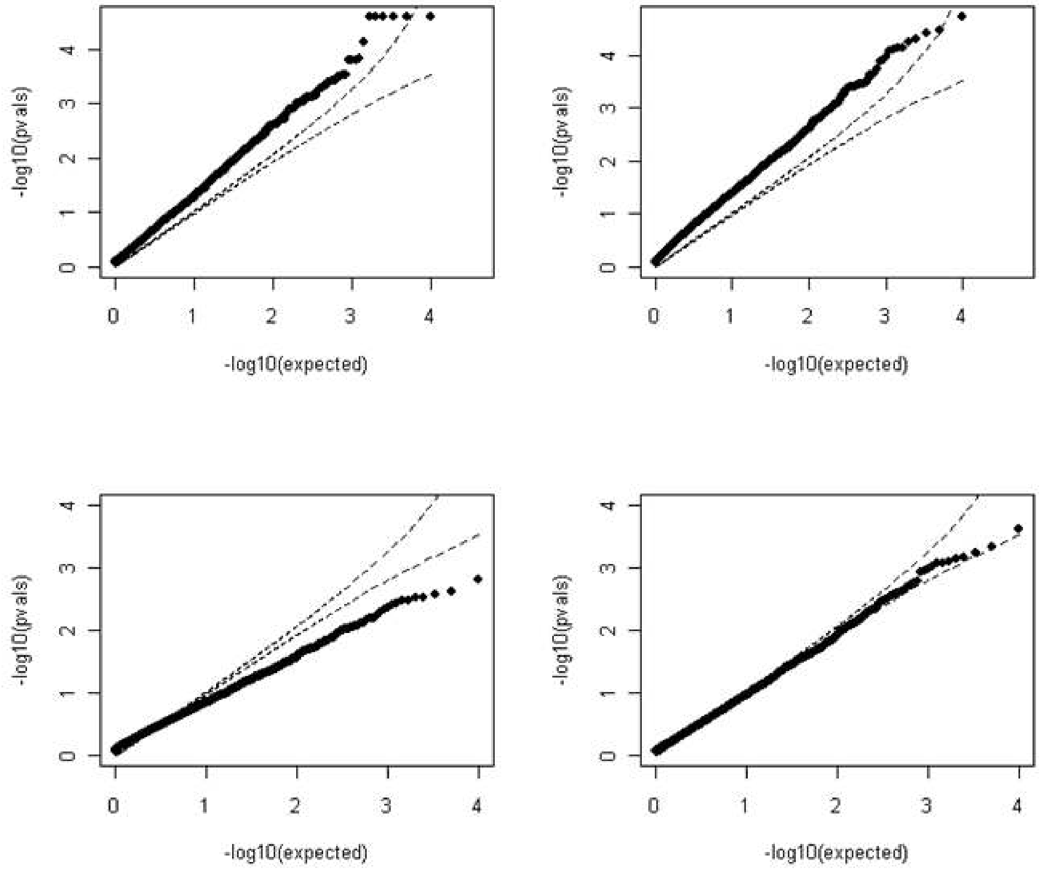
Kendall, M. G. & Stuart, A. (1969). *The Advanced Theory of Statisticsm, 3nd Edition.*



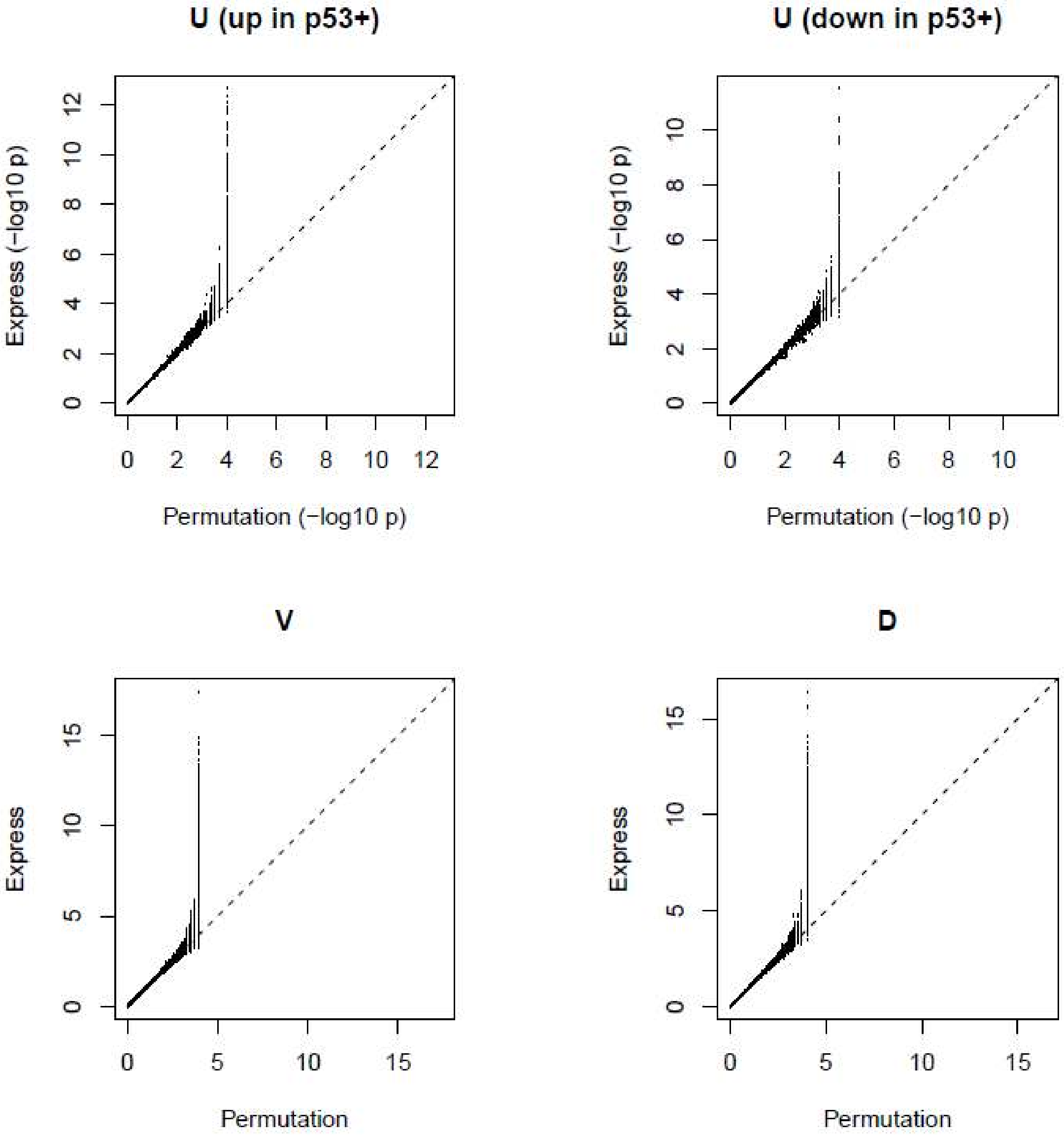
Supplementary Figure 1: p-values for V in the presence of two covariates, with 10,000 simulations for the saliva data, KEGG:00150, with one covariate correlated with both $\mathbf{X}$ and a dichotomous $\mathbf{y}$. Upper left: the result of failing to adjust for the covariates. Upper right: the result of adjusting for the covariates by residualizing X and y, but failing to account for that adjustment in computing p-values. Lower left: the result of attempting adjust Vz by a scaling factor, which fails to account properly for the variance. Lower left: the correct approach of considering the effective sample size to be reduced by the number of covariates (here, p=2).



Supplementary Figure 2: p-values for V in the presence of two covariates, with 10,000 simulations for the saliva data, KEGG:00150, with one covariate correlated with both $\mathbf{X}$ and continuous $\mathbf{y}$. The different panels follow the adjustments described above for Supplementary Figure 1.



Supplementary Figure 3: p-values for V in the presence of a covariate, with 10,000 simulations for the saliva data, KEGG:00150, with the covariate correlated with both $\mathbf{X}$ and a dichotomous $\mathbf{y}$. Upper left: the result of failing to adjust for the covariate. Upper right: the result of adjusting for the covariate by residualizing $\mathbf{X}$ and $\mathbf{y}$, but failing to account for that adjustment in computing p-values.



Supplementary Figure 4: *p*-value comparisons between *safe* (10,000 permutations) and *safe-express* on the breast cancer data. Each point corresponds to a single GO/KEGG category. Note that the permutation *p*-values are limited to 10^{-4} as a minimum value, while many safe-express values show much greater true significance. The variation in the scatterplot largely results from sampling variability due to permutation.