Larson and Owen *BMC Bioinformatics* (2015) 16:132
DOI 10.1186/s12859-015-0571-7

METHODOLOGY ARTICLE Open Access

Moment based gene set tests

Jessica L Larson[1,2*] and Art B Owen[3]

Abstract

Background: Permutation-based gene set tests are standard approaches for testing relationships between collections of related genes and an outcome of interest in high throughput expression analyses. Using M random permutations, one can attain p-values as small as $1/(M+1)$. When many gene sets are tested, we need smaller p-values, hence larger M, to achieve significance while accounting for the number of simultaneous tests being made. As a result, the number of permutations to be done rises along with the cost per permutation. To reduce this cost, we seek parametric approximations to the permutation distributions for gene set tests.

Results: We study two gene set methods based on sums and sums of squared correlations. The statistics we study are among the best performers in the extensive simulation of 261 gene set methods by Ackermann and Strimmer in 2009. Our approach calculates exact relevant moments of these statistics and uses them to fit parametric distributions. The computational cost of our algorithm for the linear case is on the order of doing $|G|$ permutations, where $|G|$ is the number of genes in set G. For the quadratic statistics, the cost is on the order of $|G|^2$ permutations which can still be orders of magnitude faster than plain permutation sampling. We applied the permutation approximation method to three public Parkinson's Disease expression datasets and discovered enriched gene sets not previously discussed. We found that the moment-based gene set enrichment p-values closely approximate the permutation method p-values at a tiny fraction of their cost. They also gave nearly identical rankings to the gene sets being compared.

Conclusions: We have developed a moment based approximation to linear and quadratic gene set test statistics' permutation distribution. This allows approximate testing to be done orders of magnitude faster than one could do by sampling permutations.
We have implemented our method as a publicly available Bioconductor package, npGSEA (www.bioconductor.org).

Keywords: GSEA, Expression analysis, Permutation tests, ROAST

Background

In a genome-wide expression study, researchers often compare the level of gene expression in thousands of genes between two treatment groups (e.g., disease, drug, phenotype, etc.). Many individual genes may trend toward differential expression, but will often fail to achieve significance. This could happen for a set of genes in a given pathway or system (a gene set). A number of significant and related genes taken together can provide strong evidence of an association between the corresponding gene set and treatment of interest. Gene set methods can improve power by looking for small, coordinated expression changes in a collection of related genes, rather than testing for large shifts in individual genes.

Additionally, single gene methods often require that all genes are independent of each other; this is not likely true in real biological systems. With known gene sets of interest, researchers can use existing biological knowledge to drive their analysis of genome-wide expression data, thereby increasing the interpretability of their results.

Mootha *et al.* [1] first introduced gene set enrichment analysis (GSEA) and calculated gene set p-values based on Kolmogorov-Smirnov statistics. Since then, there have been many methodological proposals for GSEA; no single one is always the best. For example, some tests are better for a large number of weakly associated genes, while others have better power for a small number of strongly associated genes [2].

One of the most important differences among gene set methods is the definition of the null hypothesis. Tian

*Correspondence: larson.jess@gmail.com
[1] Department of Bioinformatics and Computational Biology, Genentech, Inc., South San Francisco, USA
[2] Currently at GenePeeks, Inc., Cambridge, USA
Full list of author information is available at the end of the article

et al. [3] and Goeman and Bühlmann [4] (among others) introduce two null hypotheses that differentiate the general approaches for gene set methods. The first measures whether a gene set is more strongly related with the outcome of interest than a comparably sized gene set. Methods of this type typically rely on randomizing the gene labels to test what is often called the *competitive* null hypothesis. This is problematic because genes are inherently correlated (especially those within a set) and permuting them does not give a rigorous test [4].

The second type of approach is used to determine whether the genes within a set associate more strongly with the outcome of interest than they would by chance, had they been independent of the outcome. Methods that test this *self-contained* null hypothesis usually judge statistical significance by randomizing the phenotype with respect to expression data and assuming that gene sets are fixed. While we acknowledge that the *competitive* hypothesis is often of interest, we focus on methods that test the *self-contained* hypothesis in this paper.

Most current GSEA methods are based on random sampling of permutations. The initial GSEA [1] and widely used JG-score [5] methods both have closed form null distributions for their enrichment statistics, Kolmogorov-Smirnov and Gaussian, respectively, under appropriate assumptions. Both papers suggest permutation to gain robustness in case their assumptions don't hold.

Lehmann and Romano [6] give a concise explanation of how permutation inference works. It is common to approximate the permutation distribution by a large Monte Carlo sample [7,8]. Monte Carlo permutation tests are simple to program and do not require parametric distributional assumptions. They also can be applied to almost any statistic we might wish to investigate. However, they are often computationally expensive, are subject to random inference, and fail to achieve continuous p-values. Each of these drawbacks is described in more depth below.

Testing many sets of genes becomes computationally expensive for two reasons. First, there are many test statistics to calculate in each permuted version of the data. Second, to allow for multiplicity adjustment, we require small nominal p-values to draw inferences about our sets, which in turn requires a large number of permutations. That is, to obtain a small adjusted p-value (e.g., via FDR, FWER, Bonferroni methods), one first needs a small enough raw p-value. In order to obtain small raw p-values, the number of permutations (M) must be large, thereby increasing computational cost. Suppose that a problem requires p-values as small as ε. Rules of thumb derived in our Discussion section show that one needs to take M between $3/\varepsilon$ and $19/\varepsilon$ to get adequate power.

Because permutations are based on a random shuffling of the data, we will usually obtain a different p-value for our set of interest each time we run our permutation analysis. That is, our inference is subject to a given random seed.

Permutations are subject to two granularity issues. As mentioned above, if we do M permutations, then the smallest possible p-value we can attain is $1/(M+1)$. We call this the *resampling granularity* problem.

There is also a *data granularity* problem. In an experiment with n observations, the smallest possible p-value is at least $1/n!$. Sometimes the attainable minimum is much larger. For instance, when the target variable Y takes only the values 1 (n_1 times) and 2 (n_2 times) then the p-value cannot be smaller than $\epsilon = 1/\binom{n_1+n_2}{n_1}$. For instance, with $n_1 = n_2 = 5$, we necessarily have $p \geqslant 1/252$. More generally, when Y has tied values, taking K distinct values n_k times each, the granularity is at least $\epsilon = \Pi_{k=1}^{K} n_k!/n!$. Rotation sampling methods such as ROAST are able to get around this data granularity problem [9], under a Gaussian assumption on the data. Increased Monte Carlo sampling with methods such as ROAST can mitigate the data granularity problem but not the resampling granularity problem.

Another aspect of the resampling granularity problem is that permutations give us no basis to distinguish between two gene sets that both have the same p-value $1/(M+1)$. There may be many such gene sets, and they may have meaningfully different effect sizes. Many current approaches address this problem by ranking significantly enriched gene sets by their corresponding test statistics. This practice only works if all test statistics have the same null distribution and correlation structure, which is not the case for many current GSEA methods. Additionally, the resulting broken ties do not have a p-value interpretation and cannot be directly used in multiple testing methods. To break ties in this way also requires the retention of both a p-value and a test statistic for inference, rather than just one value.

Because of each of these limitations of permutation testing, there is a need for an alternative to sampling permutations for gene set testing. The methods we present below are moment based approximations to the distribution of some gene set test statistics. We specifically target settings where there are no outliers, and where it is extremely expensive or even infeasible to do all possible permutations or to do the desired multiple of $1/\varepsilon$ permutations. In our view, that range starts where the number of distinct permutations is about 100,000, which corresponds to binary Y with about 10 observations in each group, or continuous Y with 9 or more values. If outliers are suspected, one could replace the genes by rank statistics. If the number of distinct permutations is much smaller than 100,000 then our software prints a warning. A small number of permutations could be exhaustively enumerated, and when the number is very small, then one would not expect a moment based approximation to be suitable.

Many different gene sets tests are possible when one combines all the choices that can be made. Recently, Ackermann and Strimmer [10] compared 261 different gene set tests, and found particularly good performance from a sum of squared single gene t-test statistics. There was also good performance for a plain sum of t statistics such as the JG-score [5]. These results were surprising because the winning test statistics are among the simplest that have been proposed. They note that the performance from the sum of squares is much better than the complicated GSEA method in [11]. In their simulations the excellent performance of those two classes of statistics extended also to statistics that merely summed correlation coefficients (or their squares). Those latter statistics are the ones that we use. We develop fast approximations to the permutation p-values for weighted sums and weighted sums of squares of correlation coefficients.

Our approximate p-values are not as computationally expensive, random, or granular as their permutation counterparts. Our proposal results in a single number on the p-value scale, suitable for use in multiple comparisons algorithms. We applied our approach to three public expression analyses. Our moment based p-values closely match those from an extensive permutation analysis. They also reveal disease-associated gene sets not previously discovered in these studies.

Results

The data

For definiteness, we present our notation using the language of gene expression experiments. Let g, h, r, and s denote individual genes and G be a set of genes. The cardinality of G is denoted $|G|$, or sometimes p. That is the same letter we use for p-value, but the usages are distinct enough that there should be no confusion. Our experiment has n subjects. The subjects may represent patients, cell cultures, or tissue samples.

The expression level for gene g in subject i is X_{gi}, and Y_i is the target variable on subject i. Y_i is often a treatment, or a phenotype such as disease. We let n_k be the number of samples in the kth treatment group for K groups; $\Sigma_{k=1}^{K} n_k = n$. We center the variables so that

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_{gi} = 0, \quad \forall g. \tag{1}$$

The X_{gi} are not necessarily raw expression values, nor are they restricted to microarray values. In addition to the centering (1) they could have been scaled to have a given mean square. The scaling factor for X_{gi} might even depend on the sample variance for some genes $h \neq g$ if we thought that shrinking the variance for gene j towards the others would yield a more stable test statistic [12]. We might equally use a quantile transformation, replacing the j'th largest of the raw X_{gi} by $\Phi^{-1}((j-1/2)/n)$ where Φ is the Gaussian cumulative distribution function. Further preprocessing may be advised to handle outliers in X or Y. We do require that the preprocessing of the X's does not depend on the Y's and vice versa.

Test statistics

Our measure of association for gene g on our target variable is

$$\hat{\beta}_g = \frac{1}{n} \sum_{i=1}^{n} X_{gi} Y_i, \tag{2}$$

the sample covariance of X_{gi} and Y_i. If both X_{gi} and Y_i are centered and standardized to have variance 1, then $\hat{\beta}_g = \hat{\rho}_g$, the sample correlation between Y and gene g. The default in our software is to scale the X_{gi} values so that $\sum_{i=1}^{n} X_{gi}^2 = n$. With this default, our p-values are unaffected by scaling of Y_i and so they are equivalent to using the correlations.

If it often recommended to scale every gene to have unit variance, although the users may not always wish to. For instance in a setting where low expression values arise from probes with very low signal to noise level, scaling the genes may have the effect of inflating the noise in those probes relative to the signal in some others.

The usual t-statistic for testing a linear relationship between these variables is $t_g \equiv \sqrt{n-2}\hat{\rho}_g/(1-\hat{\rho}_g^2)^{1/2}$. A Taylor approximation to fourth order yields

$$t_g \doteq \sqrt{n-2}\left(\hat{\rho}_g + \frac{1}{2}\hat{\rho}_g^3\right) \tag{3}$$

with an error of order $\hat{\rho}_g^5$. Gene-set tests are of most use when each individual $|\hat{\rho}_g|$ is small. In such cases t_g is very nearly a constant multiple of $\hat{\rho}_g$ and we expect permutation analyses using t-statistics to be very similar to those using correlations.

For reasons of power and interpretability, we apply gene set testing methods instead of just testing individual genes. Linear and quadratic test statistics have been found to be among the best performers for gene set enrichment analyses [10]; we thus consider two statistics for our approach:

$$\widehat{T}_{G,w} = \sum_{g \in G} w_g \hat{\beta}_g \quad \text{and} \quad \widehat{C}_{G,w} = \sum_{g \in G} w_g \hat{\beta}_g^2.$$

In this paper our null hypothesis is that Y is independent of $(X_g; g \in G)$. We test this null by formulating a statistic that is sensitive to the sort of departure we think is likely, as measured by either $\widehat{T}_{G,w}$ or $\widehat{C}_{G,w}$. If it were feasible, we would use the permutation distribution of the observed test statistic to get a p-value, but to save computation we develop moment approximations instead.

When all $w_g = 1/|G|$, then $\widehat{T}_{G,w}$ reduces to the average over $g \in G$ of the correlation between X_g, when the data are standardized. Such a test statistic will be sensitive to gene sets in which the non-null genes have correlations of the same sign with Y. If we have a prior expectation that some subset of G contains genes that move in opposite directions from the others in response to changes in Y, then we may choose positive w_g for those genes and negative w_g for the rest. Similarly if some subset of the genes in G are more important to the analyst, then those genes can be given larger absolute values of w_g. The moment approximations work with general w_g.

The statistic $\widehat{T}_{G,w}$ can approximate the JG score [5]. The JG score is

$$\frac{1}{\sqrt{|G|}}\sum_{g\in G} t_g \doteq \frac{\sqrt{n-2}}{\sqrt{|G|}}\sum_{g\in G}\hat{\rho}_g = \frac{\sqrt{n-2}}{\sqrt{\mathrm{sd}(Y)|G|}}\sum_{g\in G}\frac{1}{\mathrm{sd}(X_g)}\hat{\beta}_g$$

where the approximation is good for small $\hat{\rho}_g$ and sd denotes standard deviation.

When X_g and Y are standardized then the statistics $\widehat{C}_G$ sums squared correlations. This statistics is useful when we expect that Y is associated with many of the genes $g \in G$ but we do not know *a priori* what signs to expect for the correlations, nor even to expect that they mostly share the same sign.

The letters T and C are mnemonics for the t and χ^2 distributions that resemble the permutation distributions of these quantities. The w_g are scalar weights. For the quadratic statistics we will suppose that $w_g \geqslant 0$. We won't need this condition to find moments of $C_{G,w}$. Any positive $\hat{\beta}_g^2$ contributes to evidence against the null hypothesis; negative weights would let strong evidence in one gene cancel evidence from another. Non-negative weights are also used to simplify our algorithm.

Although linear and quadratic test statistics are fairly restricted, they do allow customization through the weights w_g, and they are very interpretable compared to more ad hoc statistics. They also performed well in [10] as we describe next.

Motivation for these test statistics

Our chosen test statistics are supported by extensive simulations of Ackermann and Strimmer [10]. They compared 261 gene set testing methods. They consider per gene test statistics, that are then transformed and finally aggregated over the gene set, in various ways. Our quadratic test statistic $\widehat{C}_{G,w}$ is one of the ones that they particularly favor. The following notes are based on the summary in their pages 6–8.

They remark that they get roughly the same answers using a t-test, a moderated t-test, or a correlation, as the per gene statistic. Table two of their paper shows this. That was a surprising result because they had anticipated that moderated t-statistics might perform better. Moderated t-statistics use more stable estimates of the standard deviation of X_{gi}, suitable for small samples. See [13,14] and [15] for moderation strategies. Ackermann and Strimmer [10] offer an explanation that the lack of benefit from moderation might be due to their simulation having sample sizes as large as 10. In our target setting, the sample sizes are on the order of 10 or more.

Our $\hat{\beta}_g$ is a sample correlation when, as usual, X_{gi} and Y_i are centered and scaled variables. They remark that squaring the per gene statistics is a 'very useful transformation'. It works best on some of their scenarios. In the exceptional cases, untransformed quantities, like our linear test statistic, are best. They report that there is some advantage to a rank transformation prior to squaring. Such a transformation is possible in our framework, upon replacing X_{gi} by their ranks and then centering and scaling those ranks.

They found the mean or a maxmean over genes to be the best ways to combine the transformed statistics. We use a sum which gives the same p-values as using the mean. Medians or Wilcoxon statistics are better than the mean in one of their scenarios (correlated genes) for purposes of testing a competitive null. But that advantage vanishes when doing permutations as we do in testing the self-contained null, which is our focus here.

Finally, our linear statistic is motivated by trying to approximate the JG statistic, which is a sum of t statistics. Ackermann and Strimmer [10] found little difference between summing correlations and summing t-statistics, and our Taylor approximation above gives a reasonable explanation for their finding.

Moment based reference distributions

When we permute the data, our sample statistics $\widehat{T}_{G,w}$ and $\widehat{C}_{G,w}$ take on new values, that we denote $\widetilde{T}_{G,w}$ and $\widetilde{C}_{G,w}$. To avoid the three main disadvantages to permutation-based analyses (cost, randomness, and granularity) discussed above, we approximate the distribution of the permuted test statistics $\widetilde{T}_{G,w}$ by Gaussians or by rescaled beta distributions. For quadratic statistics $\widetilde{C}_{G,w}$ we use a distribution of the form $\sigma^2\chi^2_{(\nu)}$ choosing σ^2 and ν to match the second and fourth moments of $\widetilde{C}_{G,w}$ under permutation. The family of scaled χ^2 distributions is the same as the family of gamma distributions.

For the Gaussian treatment of $\widetilde{T}_{G,w}$ we find $\sigma^2 = \mathrm{var}\left(\widetilde{T}_{G,w}\right)$ under permutation using Eq. 8 of our Methods section and then report the p-value

$$p = \Pr\left(\mathcal{N}\left(0,\sigma^2\right) \leqslant \widehat{T}_{G,w}\right),$$

where $\widehat{T}_{G,w}$ is the observed value of the linear statistic. The above is a left tail p-value. Two-tailed and right-tailed p values are analogous.

For the linear test statistic, a scaled beta distribution provides a useful alternative to the normal distribution.

We use a scaled beta distribution, of the form $A + (B - A)\text{beta}(\alpha, \beta)$. It allows us to match four parameters of the permutation distribution (min, max, mean and variance) instead of just two as in the normal distribution. The $\text{beta}(\alpha, \beta)$ distribution has a continuous density function on $0 < x < 1$ for $\alpha, \beta > 0$. We choose A, B, α and β by matching the upper and lower limits of $\widetilde{T}_{G,w}$, as well as its mean and variance. Using Eq. 8 from our Methods section we have

$$\begin{aligned} A &= \min_{\pi} \frac{1}{n} \sum_{i=1}^{n} \sum_{g \in G} w_g X_{gi} Y_{\pi(i)}, \qquad (4) \\ B &= \max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \sum_{g \in G} w_g X_{gi} Y_{\pi(i)}, \\ \alpha &= \frac{A}{B - A} \left(\frac{AB}{\text{var}(\widetilde{T}_{G,w})} + 1 \right), \quad \text{and} \\ \beta &= \frac{-B}{B - A} \left(\frac{AB}{\text{var}(\widetilde{T}_{G,w})} + 1 \right). \end{aligned}$$

The observed left-tailed p-value is

$$p = \Pr \left(\text{beta}(\alpha, \beta) \leqslant \frac{\widehat{T}_{G,w} - A}{B - A} \right).$$

It is easy to find the permutations that maximize and minimize $\widetilde{T}_{G,w}$ by sorting the X and Y values appropriately as described in our Methods. The result has $A < 0 < B$. For the beta distribution to have valid parameters we must have $\sigma^2 < -AB$. From the inequality of Bhatia and Davis [16], we know that $\sigma^2 \leqslant -AB$. There are in fact degenerate cases with $\sigma^2 = -AB$, but in these cases $\widetilde{T}_{G,w}$ only takes one or two distinct values under permutation, and those cases are not of practical interest.

Like us, Zhou *et al.* [17] have used a beta distribution to approximate a permutation. They used the first 4 moments of a Pearson curve for their approach. Fitting by moments in the Pearson family, it is possible to get a beta distribution whose support set (A, B) does not even include the observed value $\widehat{T}_{G,w}$. That is, $\widehat{T}_{G,w}$ is even more extreme than it would have to be to get $p = 0$; it is almost like getting $p < 0$. We chose (A, B) based on the upper and lower limits of $\widetilde{T}_{G,w}$ to prevent our observed test statistic from falling outside the range of possible values of our reference distribution (Methods).

Our Beta approximation has the possibility of returning a p-value of 0 if the observed test statistic equals the most extreme possible value. A principled alternative that avoids returning 0 is to replace the left sided p_L-value by

$$\widetilde{p}_L = \epsilon + (1 - 2\epsilon) p_L$$

where ϵ is the smallest possible permutation p-value. The corresponding right and central p-values are $\widetilde{p}_R = 1 - \widetilde{p}_L$ and $\widetilde{p}_C = 2 \min (\widetilde{p}_L, \widetilde{p}_R)$. When X has a continuous distribution and Y takes K distinct values $n_1, \ldots, n_K$ times (due to ties) then the granularity is $\epsilon = \Pi_{k=1}^{K} n_k! / n!$.

For the quadratic test statistic $\widetilde{C}_{G,w}$ we use a $\sigma^2 \chi^2_{(\nu)}$ reference distribution reporting the two-tailed p-value $\Pr \left(\sigma^2 \chi^2_{(\nu)} \geqslant \widehat{C}_{G,w} \right)$ after matching the first and second moments of $\sigma^2 \chi^2_{(\nu)}$ to $\mathbb{E} \left(\widetilde{C}_{G,w} \right)$ and $\mathbb{E} \left(\widetilde{C}^2_{G,w} \right)$ respectively. The parameter values are

$$\nu = 2 \frac{\mathbb{E} \left(\widetilde{C}_{G,w} \right)^2}{\text{var} \left(\widetilde{C}_{G,w} \right)} \quad \text{and} \quad \sigma^2 = \frac{\mathbb{E} \left(\widetilde{C}_{G,w} \right)}{\nu} = \frac{\text{var} \left(\widetilde{C}_{G,w} \right)}{2\mathbb{E} \left(\widetilde{C}_{G,w} \right)}.$$

Our formulas for $\mathbb{E} \left(\widetilde{C}_{G,w} \right)$ and $\mathbb{E} \left(\widetilde{C}^2_{G,w} \right)$ under permutation are given in Eq. 5 of our Methods. Those formulas use $\mathbb{E} \left(\widetilde{\beta}^2_g \right)$ and $\text{cov} \left(\widetilde{\beta}^2_g, \widetilde{\beta}^2_h \right)$ which we give in Corollaries 1 and 2 of our Methods.

Another alternative to permutations is rotation sampling. We have also shown in our Methods section that some of the moments of our test statistics are equal to rotation moments of those test statistics. The rotation-based values for $\mathbb{E} \left(\widetilde{T}_{G,w} \right)$, $\mathbb{E} \left(\widetilde{C}_{G,w} \right)$ and $\text{var} \left(\widetilde{T}_{G,w} \right)$ are same as for permutations; the variance of $\widetilde{C}_{G,w}$ is dependent upon the choice of rotation contrast matrix.

All of our reference distributions are continuous and the χ^2 and Gaussian ones are unbounded; hence they avoid the granularity problem of permutation testing. We have prepared a publicly available Bioconductor [18] package, npGSEA, which implements our algorithm and calculates the corresponding statistics discussed in this section.

Parkinson's Disease

We illustrate our method using publicly available data from three expression studies in Parkinson's Disease (PD) patients (Table 1) [19-21]. All three experiments contain genome wide expression values measured via a microarray experiment. The values we use were normalized so that every gene had unit variance. PD is a common neurodegenerative disease; clinical symptoms often include rigidity, resting tremor and gait instability [22]. Pathologically, PD is characterized by neuronal-loss in the substantia nigra and the presence of α-synuclein protein aggregates in neurons [22].

Visualizing permutation distributions

Using a selected set from the Broad Institute's mSigDB v3.1 [23] and the presence of PD as a response variable

Table 1 Three data sets used for non-permutation GSEA

Reference	Tissue	# Affected	# Controls
Moran	Substantia nigra	29	14
Zhang	Substantia nigra	18	11
Scherzer	Blood	47	21

from the Zhang *et al.* [20] dataset, we visualized both permutation distributions and our approximation of these distributions (Figure 1). As discussed above, we use a linear test statistic, $\widehat{T}_{G,w} = \sum_{g\in G} \hat\beta_g$, and a quadratic test statistic, $\widehat{C}_{G,w} = \sum_{g\in G} \hat\beta_g^2$, where $\hat\beta_g$ is a sample covariance between gene expression and, in this case, disease status. Figure 1 shows these two test statistics with a histogram of 99,999 recomputations of those statistics for permutations of treatment status versus gene expression for a steroid signaling pathway gene set from mSigDB. It is possible for histograms of permuted test statistics to be very complicated, but in practice, they often resemble familiar parametric distributions, as in Figure 1.

Using the fitted normal distribution to determine the rarity of the observed gene set statistic results in a two-tailed p-value of 0.0604 for the linear statistic while permutations yield $p = 0.0595$. A fitted $\sigma^2\chi^2_{(\nu)}$ distribution results in $p = 0.0425$ for the sum of squares gene set statistic, while permutations yield $p = 0.0458$. The histogram for the sum of squared statistics has a somewhat sharper peak than its moment approximation. The p-values are nevertheless quite close; they are based on tail probabilities not the density itself.

Moment-based p-values tightly correlated with permutation p-values

We compared our non-permutation p-values to p-values for linear and quadratic statistics for the 6,303 gene sets from mSigDB's curated gene sets and Gene Ontology (GO) [24] gene sets collections (v3.1). One gene set was removed because it contained only one gene in our experiments. The average size of these gene sets is 79.40 genes. For our gold standard we ran 999,999 permutations of the linear statistic and 499,999 permutations of the quadratic statistic. For all of our permutations, we first calculated the observed test statistic for each of the 6,303 gene sets and then permuted the Y_i's M times to obtain 6,303 × M permuted test statistics. We next compared the precomputed test statistic vector to our matrix of permuted test statistics.

For each set, we computed left-sided p-values, p_L, for the linear statistic and two-sided p-values, p_Q, for the quadratic statistic using these permutations (Methods). We also computed the normal and beta approximations of p_L with our method. (Figure 2, left two panels). We converted these one-sided p-values to two-sided p-values via $p = 2\min(p_L, 1 - p_L)$. For very small p-values ($< 10^{-3}$), the beta and normal approximations sandwich the permutation values. At these values, the normal method is slightly conservative, while the beta approach is slightly anti-conservative. At larger p-values, the approximation-based values are almost identical to the permutation p-values.

The beta p-values can be quite a bit smaller than their permutation counterparts. Comparing two-tailed versions, we find that the beta approximate p-value is as much as 2.2-fold smaller for the Scherzer *et al.* [21] data

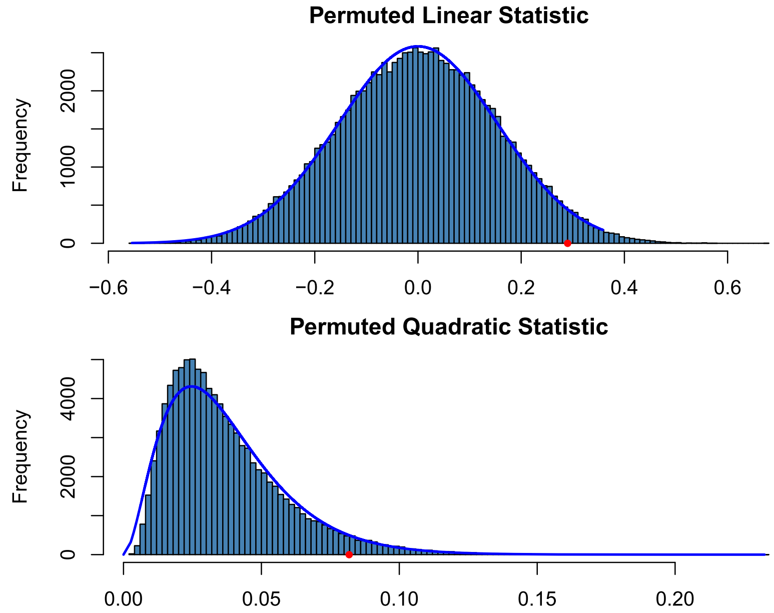

Figure 1 Distributions of permuted statistics resemble known probability densities. Top panel shows a permutation histogram for a linear test statistic for the steroid hormone signaling pathway gene set as described in the text. The bottom panel shows a quadratic test statistic. Solid red dots indicate the observed values and curves indicate parametric fits, based on normal and χ^2 distributions.

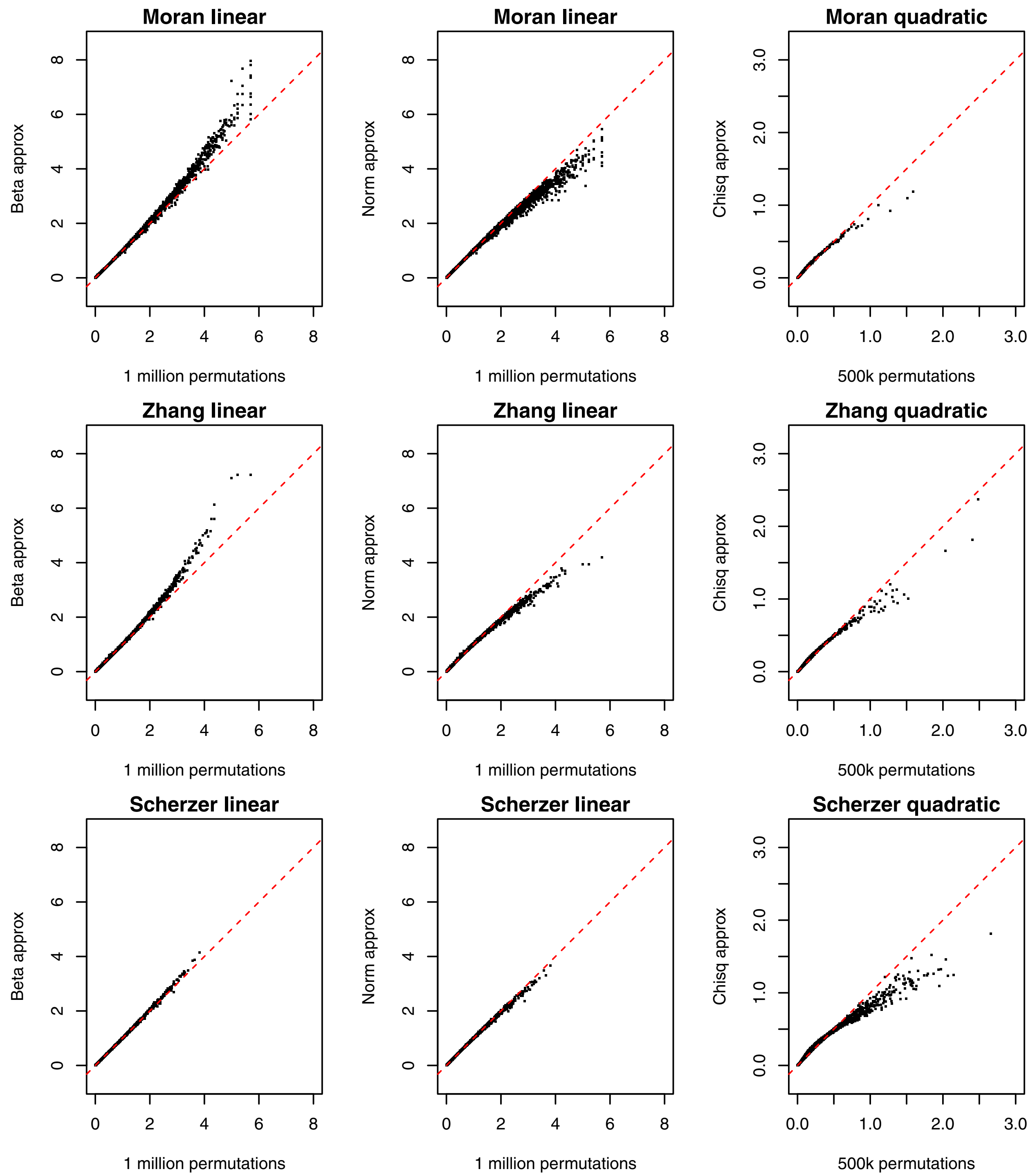

Figure 2 Permutation and moment-based *p*-values are tightly correlated. Permutation *p*-values (x-axis) versus moment-based *p*-values (y-axis) for 6,303 gene sets. The left two column represents results for a linear test statistic versus the beta and Gaussian approximations; the right-most column represents results for the sum of squares statistic versus the χ^2 approximation. Data come from three genome-wide expression studies. We applied the transformation $-\log_{10}(p)$ to stretch the lower range of these distributions for a more informative visual. Red dotted lines represent the line $y = x$.

set, 155-fold smaller for the Zhang *et al.* [20] data set, and almost 21,000-fold smaller for the Moran *et al.* [19] data set.

The very extreme ratio for the Moran data merits further investigation. It arose for a gene set in which the original data is more extreme than all 999,999 permuted versions. There were 16 gene sets where that happened. The sample of permutations does not distinguish among them; they all get a two-tailed p-value of 2×10^{-6}. The smallest beta approximate p-value is about 10^{-10}. To have sufficient power to verify such a p-value would require an extremely large number of permutations.

It is not too onerous to consider 16 tied gene sets. But a more reasonable number of permutations $M = 999$ leads to 555 gene sets tied at the most significant possible level and even $M = 9999$ leaves a tie among 186 of them.

For our quadratic test statistic, we fit our moment based $\sigma^2\chi^2_{(\nu)}$ approximation and computed two-sided p-values across all sets (Figure 2, right panel). We see that the smallest χ^2 non-permutation p-values are slightly conservative. This may reflect the boundedness of the permutation distribution combined with the unbounded right tail of the χ^2 distribution.

In each of the three experiments, there is a tight correlation between the permutation-based p-values of all sets and both of our moment-based methods (Table 2). Close rankings are important as one of the main tasks of gene set analysis is to order the gene sets so that followup investigations can be prioritized. The beta and normal approximations are almost identical. Our beta approximations are slightly closer to the gold standard than the normal approximations, but not by a practically important amount. The beta approximation has shorter tails than the Gaussian approximation. It yielded p-values somewhat smaller than permutations did, while the Gaussian approximation yielded p-values somewhat larger than the permutations did. The χ^2 approximations also reproduce the ranking of the gold standard quite well, though not as well as the normal and beta approximations to the linear statistic.

Moment-based p-values are computationally inexpensive

For these data sets and 6,303 gene sets, both of the linear statistics, which have more or less the same rank-ordering of p-values as 999,999 permutations, could be approximated in about the amount of time it takes to compute 100 permutations (Table 3, top block). This is very close to our estimated cost of $|G| \doteq 80$ permutations. While this is a close match, we remark that the time to do M permutations is nearly an affine function $a + bM$ with positive intercept a. At such small M the overhead costs dominated the total cost making the per permutation costs hard to resolve. The beta approximation was slightly slower than the Gaussian one because it involves the sorting of the data.

Table 2 Spearman correlations between gold standard (999,999 and 499,999 permutations for linear and quadratic statistics) and approximation p-values

Reference	Normal p_L	Beta p_L	Normal p_C	Beta p_C	Chisq p_Q
Moran	0.99991	0.99997	0.99973	0.99991	0.978
Zhang	0.99996	0.99997	0.99983	0.99991	0.990
Scherzer	0.99998	0.99999	0.99991	0.99997	0.994

p_{L} and p_C represent results for one and two-tailed linear test statistics, respectively. Chisq p_Q represents results for the sum of squares analysis.

Table 3 Time in seconds for p-value calculations for 6,303 gene sets in three genome-wide expression studies

Method	Moran	Zhang	Scherzer
$M = 100$	31.03	29.84	34.71
$M = 500$	31.95	32.49	35.54
$M = 1{,}000{,}000$	5010.17	4434.77	3933.15
Normal	29.74	27.00	34.66
Beta	30.79	31.88	37.89
$M = 30{,}000$	9146.27	7217.59	11808.02
$M = 40{,}000$	12256.54	9636.06	16545.60
$M = 50{,}000$	16833.08	12564.06	21480.80
$M = 500{,}000$	149588.37	129667.73	187067.91
χ^2	11020.62	10600.82	12677.15

Linear statistic results with $M = 100$, $M = 500$, and $M = 1{,}000{,}000$ permutations, and the normal and beta approximations are in the top block. Timings for the quadratic statistic with $M = 30{,}000$, $M = 40{,}000$, $M = 50{,}000$, and $M = 500{,}000$ permutations, and the χ^2 approximation are presented in the bottom block.

The χ^2 approximation to the quadratic statistic has a computational cost about as much as 35,000 to 45,000 permutations, yet has a similar rank-ordering of p-values from 499,999 permutations (Table 3, bottom block). For the quadratic statistic we expected our algorithm to cost as much as doing a number of permutations equal to a small multiple of the mean square gene set size. It cost about as much as 35,000 to 45,000 permutations while the mean square set size was 27,171.

Discovery of several gene sets associated with PD

After applying our permutation approximation methods to each dataset in 6,303 mSigDB gene sets, we found many significantly enriched gene sets, even after correcting for multiple testing with the Benjamini and Hochberg method [25] (two-sided adjusted p-value $<$ 0.05). The most significantly enriched sets are associated with metabolism and mitochondrial function, neuronal transmitters and serotonin, epigenetic modifications, and the transcription factor FOXP3 (Additional file 1: Table S1). Each of these categories has some previously discovered association with PD, although not through traditional gene set methods (metabolism and mitochondrial function [22]; neuronal transmitters and serotonin [26]; epigenetic modifications [27]; FOXP3 [28]). Through our new gene set enrichment method, we discovered a relationship between the expression of these gene sets and PD.

Discussion

Gene set methods are able to pool weak single gene signals over a set of genes to get a stronger inference. These methods and their corresponding permutation-based inferences are a staple of high throughput methods in genomics. Because an experiment for this purpose may have a few to hundreds of microarrays or RNA-seq samples, permutation can be computationally costly, and yet still result in granular p-values. In this paper, we introduce an approximate gene set method, which performs similarly to permutation methods, in a fraction of the computation time and which generates continuous p-values.

Permutation methods have some valuable properties that our approach does not share. Permutation inferences are exact at p-values that are a multiple of their underlying granularity. But typical modern gene set problems require finer resolution than permutation methods' granularity allows, because of the large number of tests being made.

The second advantage of permutations is that they apply to arbitrarily complicated statistics. In our view, many of those complicated statistics are much harder to interpret and are less intuitive than the plain sum and sum of squared statistics we present. Others have observed that simple linear and squared statistics outperform more complex approaches [10]. Our method allows for the weighting of coefficients in our statistics, granting users access to additional useful and interpretable patterns.

Because of the disadvantages discussed above, there has long been interest in finding approximations to permutation tests. Eden and Yates [7] noticed that the permutation distribution closely matched a parametric distribution that one would get running an F-test on the same data. It has also been known since the 1940s that the permutation distribution of the linear test is asymptotically normal as n increases [29].

When a problem requires p-values as small as ε then a Monte Carlo approach requires a number of sample permutations in the range of $3/\varepsilon$ to $19/\varepsilon$. The derivation is as follows. Suppose that we do $M = k/\varepsilon - 1$ permutations. We can then claim a p-value of ε or smaller if $k - 1$ or fewer sampled statistics exceed the observed value. With the true p-value (from enumeration) denoted by p, our power is then $\Pr(\text{Bin}(M, p) \leqslant k - 1)$. We suppose that the goal is to attain a p-value as small as ε with 80% power for p not much smaller than ε. For illustration, taking $\varepsilon = 10^{-6}$ with $p = 0.8\varepsilon$ and requiring power at least 80%, means that we require $k \geqslant 19$. The threshold is not sensitive to ε. The value $k = 19$ is required for $\varepsilon = 10^{-r}, p = 0.8\varepsilon$ and integers $r = 2, 3, \ldots, 40$. If we only want 80% power in the event that $p = 0.5\varepsilon$, then $k = 3$ suffices.

It may easily happen that the necessary number $M = k/\varepsilon - 1$ of permutations is onerous or even completely infeasible to do. In that case our moment based approximation provides a low cost substitute. The main limitation of our method is that we rely on a parametric approximation to the permutation distribution of our test statistic. An alternative is to employ a parametric model such as the Gaussian for X_{gi}. Unfortunately, parametric models are also inexact due to lack of fit. This applies to ROAST [9] which assumes Gaussian data. The root of the problem is the non-existence of nonparametric confidence intervals for the mean [30]. In the case of npGSEA, one can do a spot check with a modest number, say $M = 10{,}000$ permutations, to check on the accuracy of the moment based p-values.

Phipson and Smyth [31] remark that sampling permutations without replacement can be more efficient than independent sampling, and even allows access to p-values somewhat smaller than $1/(M + 1)$ especially when the number of distinct permutation values is not very large. In our target settings though, the number of distinct permutation values becomes combinatorially large, and the bookkeeping to handle sampling without replacement is cumbersome.

Knijnenburg *et al.* [32] approach the granularity issue by taking a random sample of permutations and fitting a generalized extreme value (GEV) distribution to the tail of their distribution. They use several thousand permutations, and report better ordering of gene sets using their fits than using ordinary randomization. Knijnenburg *et al.* [32] report that the observed test statistic may be larger than the maximum of their fitted GEV distribution. They find that the problem is reduced (though perhaps not eliminated) by working with either the cube or the fifth power of the test statistic.

Conclusions

We have developed a new and intuitive method for gene set enrichment analysis that is computationally inexpensive, and avoids the resampling granularity issue. A Gaussian, beta, or χ^2 approximation gives a principled way to break ties among genes or gene sets whose test statistics are larger than any seen in the M permutations. We applied our moment based approximations to three human Parkinson's Disease data sets and discovered the enrichment of several gene sets in this disease, none of which were mentioned in the original publications.

Methods

Permutation procedure

A permutation of $\{1, 2, \ldots, n\}$ is a reordering of $\{1, 2, \ldots, n\}$. There are $n!$ permutations. We call π a *uniform random permutation* of $\{1, 2, \ldots, n\}$ if it equals each distinct permutation with probability $1/n!$.

In a permutation analysis, we replace Y_i by $\widetilde{Y}_i$ where $\widetilde{Y}_i = Y_{\pi(i)}$ for $i = 1, \dots, n$. Then $\widetilde{\beta}_g = (1/n)\sum_{i=1}^n X_{gi}\widetilde{Y}_i$, and when $\widetilde{Y}$ is substituted for Y, $\widehat{T}_{G,w}$ becomes $\widetilde{T}_{G,w}$ and $\widehat{C}_{G,w}$ becomes $\widetilde{C}_{G,w}$.

The n! different permutations form a reference distribution from which we can compute p-values. There are often so many possible permutations that we cannot calculate or use all of them. Instead, we independently sample uniform random permutations M times, getting statistics $\widetilde{C}_m = \widetilde{C}_{G,w,m}$, and similarly $\widetilde{T}_m$, for $m = 1, \dots, M$. We then compute p-values by comparing our observed statistics to our permutation distribution:

$$p_Q = \frac{\#\left\{\widetilde{C}_m \geqslant \widehat{C}\right\} + 1}{M+1} \qquad p_C = \frac{\#\left\{|\widetilde{T}_m| \geqslant |\widehat{T}|\right\} + 1}{M+1}$$

$$p_L = \frac{\#\left\{\widetilde{T}_m \leqslant \widehat{T}\right\} + 1}{M+1}, \quad \text{or} \quad p_R = \frac{\#\left\{\widetilde{T}_m \geqslant \widehat{T}\right\} + 1}{M+1},$$

where p_Q and p_C are p-values for two-sided inferences on the quadratic and linear statistic, respectively, and p_L (left) and p_R (right) are for one-sided inferences based on the linear statistic. We use the mnemonic C in p_C to denote the central (or two-sided) p-value, which corresponds to a central confidence interval. The +1 in numerator and denominator of the p-values corresponds to counting the sample test statistic as one of the permutations. That is, we automatically include an identity permutation. After adding 1, the permutation distribution of the p-value is uniform on $\{1/(M+1), 2/(M+1), \dots, 1\}$.

Permutation moments of test statistics

Under permutation, $\mathbb{E}\left(\widetilde{Y}_i\right) = 0$ by symmetry, and so $\mathbb{E}\left(\widetilde{\beta}_g\right) = 0$ too. We easily find that,

$$\begin{aligned}
\mathbb{E}\left(\widetilde{T}_{G,w}\right) &= 0,\\
\mathrm{var}\left(\widetilde{T}_{G,w}\right) &= \sum_{g\in G}\sum_{h\in G} w_g w_h \mathrm{cov}\left(\widetilde{\beta}_g, \widetilde{\beta}_h\right)\\
\mathbb{E}\left(\widetilde{C}_{G,w}\right) &= \sum_{g\in G} w_g \mathbb{E}\left(\widetilde{\beta}_g^2\right), \quad \text{and} \qquad (5)\\
\mathrm{var}\left(\widetilde{C}_{G,w}\right) &= \sum_{g\in G}\sum_{h\in G} w_g w_h \mathrm{cov}\left(\widetilde{\beta}_g^2, \widetilde{\beta}_h^2\right).
\end{aligned}$$

The means, variances and covariances in (5) are taken with respect to the random permutations with the data X and Y held fixed. We adopt the convention that moments of permuted quantities are taken with respect to the permutation and are conditional on the X's and Y's. This avoids cumbersome expressions like $\mathbb{E}\left(\widetilde{\beta}_g^2 \mid X_{gi}, Y_i, g \in G\right)$.

We will need the following even moments of X and Y:

$$\begin{aligned}
\mu_2 &= \frac{1}{n}\sum_{i=1}^n Y_i^2, \quad \mu_4 = \frac{1}{n}\sum_{i=1}^n Y_i^4,\\
\bar{X}_{gh} &= \frac{1}{n}\sum_{i=1}^n X_{gi}X_{hi}, \quad \text{and}\\
\bar{X}_{ghrs} &= \frac{1}{n}\sum_{i=1}^n X_{gi}X_{hi}X_{ri}X_{si}
\end{aligned}$$

for $g, h, r, s \in G$. Although our derivations involve $O(p^4)$ different moments when the gene set G has p genes, our computations do not require all of those moments.

Lemma 1. *For an experiment with $n \geqslant 2$ including genes g and h,*

$$\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\right) = \frac{\mu_2\bar{X}_{gh}}{n-1}.$$

Proof. This appears in [33] but we prove it here to keep the paper self-contained. First

$$n^2\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\right) = \sum_i\sum_{i'} X_{gi}X_{hi'}\mathbb{E}\left(\widetilde{Y}_i\widetilde{Y}_{i'}\right)$$

Recall that $\mu_2 = \frac{1}{n}\sum_{i=1}^n Y_i^2$. Then

$$\mathbb{E}\left(\widetilde{Y}_i\widetilde{Y}_{i'}\right) = \begin{cases} \mu_2, & i' = i\\ -\dfrac{1}{n-1}\mu_2, & i' \neq i\end{cases}$$

and so

$$\begin{aligned}
n^2\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\right) &= \sum_i\sum_{i'} X_{gi}X_{hi'}\mathbb{E}\left(\widetilde{Y}_i\widetilde{Y}_{i'}\right)\\
&= \mu_2\sum_i\sum_{i'} X_{gi}X_{hi'}\left(1_{i=i'} - \frac{1}{n-1}1_{i\neq i'}\right)\\
&= \mu_2\sum_i\sum_{i'} X_{gi}X_{hi'}\left(\frac{n}{n-1}1_{i=i'} - \frac{1}{n-1}\right)\\
&= \frac{n}{n-1}\mu_2\sum_i X_{gi}X_{hi}\\
&\equiv \frac{n^2}{n-1}\mu_2\bar{X}_{gh},
\end{aligned}$$

proving Lemma 1. □

Corollary 1. *For an experiment with $n \geqslant 2$ including genes g and h,*

$$cov\left(\widetilde{\beta}_g, \widetilde{\beta}_h\right) = \mu_2\bar{X}_{gh}/(n-1).$$

Proof. This follows from Lemma 1 because $\mathbb{E}\left(\widetilde{\beta}_g\right) = 0$. □

From Corollary 1, we see that the correlation between permuted test statistics $\widetilde{\beta}_g$ and $\widetilde{\beta}_h$ is simply the correlation between expression values for genes g and h.

Lemma 2. *For an experiment with $n \geqslant 4$ including genes g, h, r, s,*

$$\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\widetilde{\beta}_r\widetilde{\beta}_s\right) = \begin{pmatrix} \mu_2^2 \\ \mu_4 \end{pmatrix}^{\mathsf{T}} A^{\mathsf{T}} B \begin{pmatrix} \bar{X}^*_{ghrs}/n^2 \\ \bar{X}_{ghrs}/n^3 \end{pmatrix}$$

where $\bar{X}^*_{ghrs} = \bar{X}_{gh}\bar{X}_{rs} + \bar{X}_{gs}\bar{X}_{hr} + \bar{X}_{gr}\bar{X}_{hs}$, with A^{T} given by

$$\begin{pmatrix} 0 & 0 & \frac{n}{n-1} & \frac{-n}{(n-1)(n-2)} & \frac{3n}{(n-1)(n-2)(n-3)} \\ 1 & \frac{-1}{n-1} & \frac{-1}{n-1} & \frac{2}{(n-1)(n-2)} & \frac{-6}{(n-1)(n-2)(n-3)} \end{pmatrix},$$

and

$$B = \begin{pmatrix} 0 & 1 \\ 0 & -4 \\ 1 & -3 \\ -2 & 12 \\ 1 & -6 \end{pmatrix}.$$

Proof. The fourth moment contains terms of the form

$$X_{gi}X_{hj}X_{rk}X_{s\ell}\mathbb{E}\left(\widetilde{Y}_i\widetilde{Y}_j\widetilde{Y}_k\widetilde{Y}_\ell\right)$$

and there are different special cases depending on which pairs of indices among i, j, k and ℓ are equal. We need the following fourth moments of Y in which all indices are distinct:

$$\begin{aligned} \mu_{4k} &= \mathbb{E}\left(\widetilde{Y}_i^4\right) \\ \mu_{3k} &= \mathbb{E}\left(\widetilde{Y}_i^3\widetilde{Y}_j\right) \\ \mu_{2p} &= \mathbb{E}\left(\widetilde{Y}_i^2\widetilde{Y}_j^2\right) \\ \mu_{1p} &= \mathbb{E}\left(\widetilde{Y}_i^2\widetilde{Y}_j\widetilde{Y}_k\right) \\ \mu_{\varnothing} &= \mathbb{E}\left(\widetilde{Y}_i\widetilde{Y}_j\widetilde{Y}_k\widetilde{Y}_\ell\right), \end{aligned}$$

and where the subscripts are mnemonics for terms four of a kind, three of a kind, two pair, one pair and nothing special.

We can express all of these moments in terms of μ_2 and $\mu_4 = (1/n)\sum_{i=1}^n Y_i^4$. Each moment is a normalized sum over distinct indices. We can write these in terms of normalized sums over all indices. Many of those terms vanish because $\sum_i Y_i = 0$.

Let $\sum^*$ represent summation over distinct indices, as in

$$\begin{aligned} \sum_{ij}{}^{*} f_{ij} &= \sum_{i=1}^n \sum_{j=1, j\neq i}^n f_{ij}, \\ \sum_{ijk}{}^{*} f_{ijk} &= \sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{k=1, k\neq i, k\neq j} f_{ijk} \end{aligned}$$

and so on. We can write these sums in terms of unrestricted sums:

$$\begin{aligned} \sum_{ij}{}^{*} f_{ij} &= \sum_{ij} f_{ij} - \sum_i f_{ii} \\ \sum_{ijk}{}^{*} f_{ijk} &= \sum_{ijk} f_{ijk} - \sum_{ij}\left(f_{iij} + f_{iji} + f_{ijj}\right) + 2\sum_i f_{iii}, \quad \text{and} \\ \sum_{ijk\ell}{}^{*} f_{ijk\ell} &= \sum_{ijk\ell} f_{ijk\ell} - \sum_{ijk}\left(f_{ijki} + f_{ijkj} + f_{ijkk} + f_{ijik} + f_{ijjk} + f_{iijk}\right) \\ &\quad + \sum_{ij}\left(2\left(f_{ijjj} + f_{ijii} + f_{iiji} + f_{iiij}\right) + f_{ijij} + f_{ijji} + f_{iijj}\right) \\ &\quad -6\sum_i f_{iiii}. \end{aligned}$$

See Gleich and Owen [34] for details.

We will use the last expression in a context where $f_{ijk\ell}$ vanishes when summed over the entire range of any one of its indices. In that case

$$\sum_{ijk\ell}{}^{*} f_{ijk\ell} = \sum_{ij}\left(f_{ijij} + f_{ijji} + f_{iijj}\right) - 6\sum_i f_{iiii}. \qquad (6)$$

We also use the notation $n^{(k)} = n(n-1)(n-2)\cdots(n-k+1)$, often called '$n$ to k factors', where k is a positive integer. Now

$$\begin{aligned} \mu_{4k} &= \frac{1}{n}\sum_{i=1}^n Y_i^4 = \mu_4, \\ \mu_{3k} &= \frac{1}{n^{(2)}}\sum_{ij}{}^{*} Y_i^3 Y_j = \frac{1}{n^{(2)}}\left(\sum_{ij} Y_i^3 Y_j - \sum_i Y_i^4\right) \\ &= -\frac{\mu_4}{n-1}, \\ \mu_{2p} &= \frac{1}{n^{(2)}}\sum_{ij}{}^{*} Y_i^2 Y_j^2 = \frac{1}{n^{(2)}}\left(\sum_{ij} Y_i^2 Y_j^2 - \sum_i Y_i^4\right) \\ &= \frac{1}{n-1}\left(n\mu_2^2 - \mu_4\right), \quad \text{and} \\ \mu_{1p} &= \frac{1}{n^{(3)}}\sum_{ijk}{}^{*} Y_i^2 Y_j Y_k \\ &= \frac{1}{n^{(3)}}\left(\sum_{ijk} Y_i^2 Y_j Y_k - \sum_{ij}\left(2Y_i^3 Y_j + Y_i^2 Y_j^2\right) + 2\sum_i Y_i^4\right) \\ &= \frac{-n\mu_2^2 + 2\mu_4}{(n-1)(n-2)}. \end{aligned}$$

Finally using (6), $n^{(4)}\mu_{\varnothing}$ equals

$$\sum_{ijk\ell}{}^{*} Y_i Y_j Y_k Y_\ell = 3\sum_{ij} Y_i^2 Y_j^2 - 6\sum_i Y_i^4 = 3n^2\mu_2^2 - 6n\mu_4$$

so that

$$\mu_{\varnothing} = \frac{1}{(n-1)(n-2)(n-3)}\left(3n\mu_2^2 - 6\mu_4\right).$$

We may summarize these results via

$$\begin{pmatrix} \mu_{4k} \\ \mu_{3k} \\ \mu_{2p} \\ \mu_{1p} \\ \mu_{\varnothing} \end{pmatrix} = A \begin{pmatrix} \mu_2^2 \\ \mu_4 \end{pmatrix},$$

where the matrix A is given in the statement of Lemma 2.

Now

$$\begin{aligned} n^4 \mathbb{E}\left(\widetilde{\beta}_g \widetilde{\beta}_h \widetilde{\beta}_r \widetilde{\beta}_s\right) &= \sum_{ijk\ell} X_{gi} X_{hj} X_{rk} X_{s\ell} \mathbb{E}(\widetilde{Y}_i \widetilde{Y}_j \widetilde{Y}_k \widetilde{Y}_\ell) \\ &= \mu_{4k} \sum_i X_{gi} X_{hi} X_{ri} X_{si} \\ &\quad + \mu_{3k} \sum_{ij}^{*} \big(X_{gi} X_{hi} X_{ri} X_{sj} + X_{gi} X_{hi} X_{rj} X_{si} \\ &\quad + X_{gi} X_{hj} X_{ri} X_{si} + X_{gj} X_{hi} X_{ri} X_{si} \big) \\ &\quad + \mu_{2p} \sum_{ij}^{*} \big(X_{gi} X_{hi} X_{rj} X_{sj} + X_{gi} X_{hj} X_{ri} X_{sj} \\ &\quad + X_{gi} X_{hj} X_{rj} X_{si} \big) \\ &\quad + \mu_{1p} \sum_{ijk}^{*} \big(X_{gi} X_{hi} X_{rj} X_{sk} + X_{gi} X_{hj} X_{ri} X_{sk} \\ &\quad + X_{gi} X_{hj} X_{rk} X_{si} + X_{gi} X_{hj} X_{rj} X_{sk} \\ &\quad + X_{gi} X_{hj} X_{rk} X_{sj} + X_{gi} X_{hj} X_{rk} X_{sk} \big) \\ &\quad + \mu_{\varnothing} \sum^{*} X_{gi} X_{hj} X_{rk} X_{s\ell}. \end{aligned}$$

Next, we write the terms of $n^4 \mathbb{E}\left(\widetilde{\beta}_g \widetilde{\beta}_h \widetilde{\beta}_r \widetilde{\beta}_s\right)$ using $\bar{X}_{ghrs}$ and similar moments.

The coefficient of μ_{4k} is $\sum_i X_{gi} X_{hi} X_{ri} X_{si} = n\bar{X}_{ghrs}$. The coefficient of μ_{3k} contains

$$\begin{aligned} \sum_{ij}^{*} X_{gi} X_{hi} X_{ri} X_{sj} &= \sum_{ij} X_{gi} X_{hi} X_{ri} X_{sj} - \sum_i X_{gi} X_{hi} X_{ri} X_{si} \\ &= -n\bar{X}_{ghrs} \end{aligned}$$

and after summing all four such terms, the coefficient is $-4n\bar{X}_{ghrs}$. The coefficient of μ_{2p} contains

$$\begin{aligned} \sum_{ij}^{*} X_{gi} X_{hi} X_{rj} X_{sj} &= \sum_{ij} X_{gi} X_{hi} X_{rj} X_{sj} - \sum_i X_{gi} X_{hi} X_{ri} X_{si} \\ &= -n\bar{X}_{ghrs} \end{aligned}$$

and accounting for all three terms yields $-3n\bar{X}_{ghrs}$.

The coefficient of μ_{1p} contains

$$\begin{aligned} \sum_{ijk}^{*} X_{gi} X_{hi} X_{rj} X_{sk} &= \sum_{ijk} X_{gi} X_{hi} X_{rj} X_{sk} - \sum_{ij} X_{gi} X_{hi} X_{ri} X_{sj} \\ &\quad - \sum_{ik} X_{gi} X_{hi} X_{rj} X_{si} - \sum_{jk} X_{gi} X_{hi} X_{rj} X_{sj} \\ &\quad + 2 \sum_i X_{gi} X_{hi} X_{ri} X_{si} \\ &= -n^2 \bar{X}_{gh} \bar{X}_{rs} + 2n\bar{X}_{ghrs}. \end{aligned}$$

Summing all 6 terms, we find that the coefficient is

$$-2n^2 \left(\bar{X}_{gh} \bar{X}_{rs} + \bar{X}_{gr} \bar{X}_{hs} + \bar{X}_{gs} \bar{X}_{hr} \right) + 12n\bar{X}_{ghrs}.$$

The coefficient of $\mu_{\varnothing}$ is, using (6),

$$\begin{aligned} \sum_{ijk\ell}^{*} X_{gi} X_{hj} X_{rk} X_{s\ell} &= \sum_{ij} \big(X_{gi} X_{hj} X_{ri} X_{sj} + X_{gi} X_{hj} X_{rj} X_{si} \\ &\quad + X_{gi} X_{hi} X_{rj} X_{sj} \big) - 6 \sum_i X_{gi} X_{hi} X_{ri} X_{si} \\ &= n^2 \big(\bar{X}_{gh} \bar{X}_{rs} + \bar{X}_{gr} \bar{X}_{hs} + \bar{X}_{gs} \bar{X}_{hr} \big) - 6n\bar{X}_{ghrs}. \end{aligned}$$

We may summarize these results via

$$\mathbb{E}\left(\widetilde{\beta}_g \widetilde{\beta}_h \widetilde{\beta}_r \widetilde{\beta}_s\right) = \begin{pmatrix} \mu_{4k} \\ \mu_{3k} \\ \mu_{2p} \\ \mu_{1p} \\ \mu_{\varnothing} \end{pmatrix}^{\mathsf{T}} B \begin{pmatrix} \bar{X}^{*}_{ghrs}/n^2 \\ \bar{X}_{ghrs}/n^3 \end{pmatrix}, \quad \text{for}$$

$$B = \begin{pmatrix} 0 & 1 \\ 0 & -4 \\ 1 & -3 \\ -2 & 12 \\ 1 & -6 \end{pmatrix},$$

where $\bar{X}^{*}_{gh,rs} = \bar{X}_{gh} \bar{X}_{rs} + \bar{X}_{gr} \bar{X}_{hs} + \bar{X}_{gs} \bar{X}_{hr}$, completing the proof of Lemma 2. □

These moment expressions have been checked by comparing the variance expression for the quadratic test statistic to that obtained by enumerating all permutations of a small data set. They match.

The expression in Lemma 2 is complicated, but it is simple to compute; we need only two moments of Y, two cross-moments of X, and the 2×2 matrix $A^{\mathsf{T}} B$. The matrix A depends on the experiment through n. Using Lemma 2 we can obtain the covariance between $\widetilde{\beta}_g^2$ and $\widetilde{\beta}_h^2$.

Corollary 2. *For an experiment with $n \geqslant 4$, and genes g, h,*

$$cov\left(\widetilde{\beta}_g^2, \widetilde{\beta}_h^2\right) = \begin{pmatrix} \mu_2^2 \\ \mu_4 \end{pmatrix}^{\mathsf{T}} A^{\mathsf{T}} B \begin{pmatrix} \bar{X}^{*}_{gghh}/n^2 \\ \bar{X}_{gghh}/n^3 \end{pmatrix} - \frac{\mu_2^2}{(n-1)^2} \bar{X}_{gg} \bar{X}_{hh},$$

where $\bar{X}^{*}_{gghh} = \bar{X}_{gg} \bar{X}_{hh} + 2\bar{X}_{gh}^2$ with A and B as given in Lemma 2.

Proof. The covariance is $\mathbb{E}\left(\widetilde{\beta}_g^2\widetilde{\beta}_h^2\right) - \mathbb{E}\left(\widetilde{\beta}_g^2\right)\mathbb{E}\left(\widetilde{\beta}_h^2\right)$. Applying Lemma 2 to the first expectation and Lemma 1 to the other two yields the result. □

Rotation moments of test statistics

Rotation sampling [35,36] provides an alternative to permutations, and is justified if either X or Y has a Gaussian distribution. It is simple to describe when $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, and simplifies further in the special case $\mu = 0$. In the latter case we can replace Y by $\widetilde{Y} = QY$ where $Q \in \mathbb{R}^{n\times n}$ is a random orthogonal matrix (independent of both X and Y), and the distribution of our test statistics is unchanged under the null hypothesis that X and Y are independent.

Rotation tests work by repeatedly sampling from the uniform distribution on random orthogonal matrices and recomputing the test statistics using $\widetilde{Y}$ instead of Y. They suffer from resampling granularity but not data granularity because Q has a continuous distribution (for $n \geqslant 2$).

To take account of centering we need to use a rotation test appropriate for $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$. Langsrud [36] does this by choosing rotation matrices that leave the population mean fixed. He rotates the data in an $n-1$ dimensional space orthogonal to the vector 1_n. To get such a rotation matrix, he first selects an orthogonal contrast matrix $W \in \mathbb{R}^{n\times(n-1)}$. This matrix satisfies $W^\mathsf{T}W = I_{n-1}$ and $W^\mathsf{T}1_n = 0_{n-1}$. Then he generates a uniform random rotation $Q^* \in \mathbb{R}^{(n-1)\times(n-1)}$ and delivers $\widetilde{Y} = QY$, where $Q = \frac{1}{n}1_n1_n^\mathsf{T} + WQ^*W^\mathsf{T}$. More generally if $Y \sim \mathcal{N}(Z\gamma, \sigma^2 I_n)$, for a linear model $Z\gamma$, Langsrud [36] shows how to rotate Y in the residual space of this model, leaving the fits unchanged.

Wu *et al.* [9] have implemented rotation sampling for microarray experiments in their method, ROAST. They speed up the sampling by generating a random vector instead of a random matrix. For some tests, permutations and rotations have the same moments, and so our approximations are approximations of rotation tests as much as of permutation tests.

Our rotation method approximation performs very similarly to the permutation method. We let $\widetilde{Y} = QY$ for $Q = (\frac{1}{n}1_n1_n^\mathsf{T} + WQ^*W^\mathsf{T})$ where Q^* is a uniform random $n-1 \times n-1$ rotation matrix and the contrast matrix $W \in \mathbb{R}^{n\times(n-1)}$ satisfies $W^\mathsf{T}1_n = 0_{n-1}$ and $W^\mathsf{T}W = I_{n-1}$ and then $\widetilde{\beta}$, $\widetilde{T}$ and $\widetilde{C}$ are defined as for permutations, substituting $\widetilde{Y}$ for Y.

The variance of the quadratic test statistic depends on *which* contrast matrix W one chooses, and so it cannot always match the permutation variance. This difference disappears asymptotically as $n \to \infty$. Our main results on rotation sampling are that the other moments match, as follows.

Lemma 3. *For an experiment with $n \geqslant 2$ including genes g and h, the moments $\mathbb{E}\left(\widetilde{\beta}_g\right)$ and $\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\right)$ are identical to their permutation counterparts, regardless of the choice for W.*

We prove Lemma 3 below. It has the following immediate consequence.

Corollary 3. *For an experiment with $n \geqslant 2$, $\mathbb{E}\left(\widetilde{T}_{G,w}\right)$, var $\left(\widetilde{T}_{G,w}\right)$ and $\mathbb{E}\left(\widetilde{C}_{G,w}\right)$ are the same whether $\widetilde{Y}$ is formed by permutation or rotation of Y.*

Proof of Lemma 3. We begin with some low order moments of orthogonal random matrices. For integers $n \geqslant k \geqslant 1$, let $V_{n,k} = \left\{Q \in \mathbb{R}^{n\times k} \mid Q^\mathsf{T}Q = I_k\right\}$, known as the Stiefel manifold. We will make use of the uniform distributions on $V_{n,k}$. There is a natural identification of $V_{n,1}$ with the unit sphere.

Let $Q \in \mathbb{R}^{n\times n}$ be a uniform random rotation matrix. This implies, among other things, that each column of Q is a uniform random point on the unit sphere in n dimensions.

By symmetry, we find that $\mathbb{E}(Q_{ij}) = 0$. Similarly $\mathbb{E}(Q_{ij}^2) = \mathbb{E}((1/n)\sum_{j=1}^n Q_{ij}^2) = 1/n$ and $\mathbb{E}(Q_{ij}Q_{rs}) = 0$ unless $i = r$ and $j = s$. Let $X_i \in \mathbb{R}^p$ where $p = |G|$ and $Y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Both X_i and Y_i are centered: $\sum_i X_i = 0$ and $\sum_i Y_i = 0$.

The sample coefficients for genes $g \in G$ are given by the vector $\hat{\beta} = (1/n)\sum_i X_iY_i \in \mathbb{R}^{|G|}$. The reference distribution is formed by sampling values of $\widetilde{\beta} = (1/n)\sum_i X_i\widetilde{Y}_i$ where $\widetilde{Y}$ is a rotated version of Y.

The rotation is one that preserves the mean of Y while rotating in the $n-1$ dimensional space of contrasts. As in [36], we let $W \in \mathbb{R}^{n\times(n-1)}$ be any fixed contrast matrix satisfying $W^\mathsf{T}W = I_{n-1}$ and $W^\mathsf{T}1_n = 0_{n-1}$. Then the rotated version of Y is

$$\widetilde{Y} = WQW^\mathsf{T}Y, \quad \text{where} \quad Q \sim \mathsf{U}(V_{n-1,n-1})$$

is a uniform random $n-1$ dimensional rotation matrix.

It is convenient to introduce centered quantities $X^c = W^\mathsf{T}X \in \mathbb{R}^{(n-1)\times p}$, $Y^c = W^\mathsf{T}Y \in \mathbb{R}^{n-1}$ and $\widetilde{Y}^c = W^\mathsf{T}\widetilde{Y} \in \mathbb{R}^{n-1}$. These sum to zero even when X, Y and $\widetilde{Y}$ do not. Their main difference from those variables is that they have $n-1$ rows, not n.

Now $\widetilde{\beta} = (1/n)X^\mathsf{T}\widetilde{Y} = (1/n)X^\mathsf{T}WQW^\mathsf{T}Y = (1/n)X^{c\mathsf{T}}QY^c$, so

$$\mathbb{E}\left(\widetilde{\beta}\right) = (1/n)X^{c\mathsf{T}}\mathbb{E}(Q)Y^{c\mathsf{T}} = 0,$$

matching the moment under permutation. For the rest of the proof, we need the covariance matrix of $\widetilde{\beta}$. Now

$$\mathbb{E}\left(\widetilde{\beta}\widetilde{\beta}^\mathsf{T}\right) = \frac{1}{n^2}X^{c\mathsf{T}}\mathbb{E}\left(Q^\mathsf{T}Y^cY^{c\mathsf{T}}Q\right)X^{c\mathsf{T}} = \frac{1}{n^2}X^{c\mathsf{T}}\mathbb{E}\left(Q^\mathsf{T}ZQ\right)X^c$$

where $Z = Y^cY^{c\mathsf{T}} \in \mathbb{R}^{(n-1)\times(n-1)}$.

The ij element of $Q^{\mathsf{T}}ZQ$ is $(Q^{\mathsf{T}}ZQ)_{ij} = \sum_{k=1}^{n-1}\sum_{\ell=1}^{n-1} Z_{k\ell} Q_{ki}Q_{\ell j}$ which has expected value

$$\sum_{k=1}^{n-1}\sum_{\ell=1}^{n-1} Z_{k\ell}1_{k=\ell}1_{i=j}/(n-1) = \frac{1_{i=j}}{n-1}\sum_{k=1}^{n-1} Z_{kk} = 1_{i=j}\frac{n}{n-1}\mu_2$$

where $\mu_2 = (1/n)\sum_{i=1}^{n} Y_i^2 = (1/n)\sum_{i=1}^{n} Y_i^{c2}$. That is

$$\mathbb{E}\left(Q^{\mathsf{T}}ZQ\right) = \frac{n\mu_2}{n-1}I_{n-1}$$

and so

$$\mathbb{E}\left(\widetilde{\beta}\widetilde{\beta}^{\mathsf{T}}\right) = \frac{\mu_2}{n(n-1)}X^{c\mathsf{T}}X^c.$$

In particular $\mathbb{E}\left(\widetilde{\beta}_g\widetilde{\beta}_h\right) = \mathbb{E}\left(\widetilde{\beta}\widetilde{\beta}^{\mathsf{T}}\right)_{gh} = \bar{X}_{gh}\mu_2/(n-1)$, matching the value under permutation. □

Fourth moments

Here we show that the variance of $\widetilde{C}_{G,w}$ in rotation sampling can depend on the specific matrix W used. We need fourth moments like $\mathbb{E}\left(\widetilde{\beta}_r^2\widetilde{\beta}_s^2\right)$. Those in turn depend on fourth moments of Q.

Anderson, Olkin and Underhill [37] give

$$\mathbb{E}\left(Q_{ij}^4\right) = \frac{3}{n(n+2)}. \qquad (7)$$

We are interested in all fourth moments $\mathbb{E}(Q_{ij}Q_{k\ell} Q_{rs}Q_{tu})$ of Q. If any of j, ℓ, s, u appears exactly once then the fourth moment is 0 by symmetry. To see this, suppose that index ℓ appears exactly once. Now define the matrix $\widetilde{Q}$ with elements

$$\widetilde{Q}_{ij} = \begin{cases} -Q_{ij} & j = \ell, \\ Q_{ij} & j \neq \ell. \end{cases}$$

If $Q \sim \mathsf{U}(V_{n,n})$ then $\widetilde{Q} \sim \mathsf{U}(V_{n,n})$ too by invariance of $\mathsf{U}(V_{n,n})$ to multiplication on the right by the orthogonal matrix diag$(1, 1, \dots, 1, -1, 1, \dots, 1)$, with a -1 in the j'th position. Then

$$\begin{aligned}\mathbb{E}(Q_{ij}Q_{k\ell}Q_{rs}Q_{tu}) &= \frac{1}{2}\mathbb{E}\left(Q_{ij}Q_{k\ell}Q_{rs}Q_{tu} + \widetilde{Q}_{ij}\widetilde{Q}_{k\ell}\widetilde{Q}_{rs}\widetilde{Q}_{tu}\right) \\ &= \frac{1}{2}\mathbb{E}\Big(Q_{ij}Q_{k\ell}Q_{rs}Q_{tu} + Q_{ij}(-Q_{k\ell})Q_{rs}Q_{tu}\Big) \\ &= 0.\end{aligned}$$

Similarly, because Q^{T} is also uniformly distributed on $V_{n,n}$ we find that if any of i, k, r, t appear exactly once the moment is zero. If one index appears exactly three times, then some other moment must appear exactly once. As a result, the only nonzero fourth moments are products of squares and pure fourth moments. Their values are given in the Lemma below.

Lemma 4. *Let $Q \sim \mathsf{U}(V_{n,n})$. Then*

$$\mathbb{E}\left(Q_{ij}^2Q_{rs}^2\right) = \begin{cases} \frac{3}{n(n+2)}, & i = r \,\&\, j = s \\ \frac{1}{n(n+2)}, & 1_{i=r} + 1_{j=s} = 1 \\ \frac{n+1}{n(n-1)(n+2)}, & i \neq r \,\&\, j \neq s. \end{cases}$$

Proof. The first case was given by [37]. For the second case, there is no loss of generality in computing $\mathbb{E}\left(Q_{11}^2Q_{21}^2\right)$. The vector $(Q_{11}, Q_{21}, \dots, Q_{n1})$ is uniformly distributed on the sphere. Given Q_{11}, the point $(Q_{21}, Q_{31}, \dots, Q_{n1})$ is uniformly distributed on the $n-1$ dimensional sphere of radius $\sqrt{1-Q_{11}^2}$. Therefore $\mathbb{E}\left(Q_{21}^2 \mid Q_{11}\right) = \left(1-Q_{11}^2\right)/(n-1)$ and so

$$\begin{aligned}\mathbb{E}\left(Q_{11}^2Q_{21}^2\right) &= \frac{1}{n-1}\mathbb{E}\left(Q_{11}^2 - Q_{11}^4\right) \\ &= \frac{1}{n-1}\left(\frac{1}{n} - \frac{3}{n(n+2)}\right) = \frac{1}{n(n+2)}.\end{aligned}$$

For the remaining case we let $\theta = \mathbb{E}(Q_{ij}^2Q_{rs}^2)$ for $i \neq r$ and $j \neq s$. Summing over n^4 combinations of indices we find that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\sum_{s=1}^{n} Q_{ij}^2Q_{rs}^2 = \left(\sum_{ij} Q_{ij}^2\right)^2 = n^2$$

by orthogonality of Q. Therefore

$$\begin{aligned}n^2 &= \mathbb{E}\left(\sum_{ij}\sum_{rs} Q_{ij}^2Q_{rs}^2\right) \\ &= n^2\mathbb{E}\left(Q_{11}^4\right) + 2n^2(n-1)\mathbb{E}(Q_{11}^2Q_{12}^2) + n^2(n-1)^2\theta.\end{aligned}$$

Solving for θ we get

$$\theta = \frac{n^2 - \frac{3n}{n+2} - \frac{2n(n-1)}{n+2}}{n^2(n-1)^2} = \frac{n+1}{n(n-1)(n+2)}.$$

□

The exact value of $\mathbb{E}\left(\widetilde{\beta}_r^2\widetilde{\beta}_s^2\right)$ is a very bulky expression. It does however include a term with a nonzero coefficient multiplied by $\sum_{i=1}^{n}(Y_i^c)^4$ times a similar quantity involving X. This fourth moment depends on the matrix W used. To see this in an example consider that for $n = 3$, we could take

$$W^{\mathsf{T}} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix}$$

Then $\sum_i (W^{\mathsf{T}}Y)_i^4 = (5/9)Y_1^4 + (5/9)Y_2^4 + (1/9)Y_3^4$. Permuting the columns of W^{T} would then change which Y_i got the small coefficient. Lemma 4 convinces us that the effect of W on ROAST vanishes for var$(\widetilde{C}_{G,w})$ as n increases. That Lemma shows that the cross moments

$\mathbb{E}\left(Q_{ij}^2 Q_{rs}^2\right)$ for $i \neq r$ or $j \neq s$, are of the same order of magnitude as $\mathbb{E}(Q_{ij}^4)$. Those moments appear in coefficients of only second moments of $W^{\mathsf{T}}Y$ and $X^{\mathsf{T}}Y$. Also there are many more of them so they dominate the cross moments $\mathbb{E}\left(\widetilde{\beta}_r^2 \widetilde{\beta}_s^2\right)$.

Computation and costs

To facilitate computation for the linear statistic, we reduce each gene set to a single pseudo-gene $X_{Gi} = \sum_{g \in G} w_g X_{gi}$ and then let

$$\bar{X}_G = \frac{1}{n}\sum_{i=1}^{n} X_{Gi} \quad \text{and} \quad \bar{X}_{GG} = \frac{1}{n}\sum_{i=1}^{n} X_{Gi}^2.$$

The weights w have been absorbed into the pseudo-gene to simplify notation. We define

$$\begin{aligned} \hat{\beta}_G &= \sum_{g \in G} w_g \hat{\beta}_g = \frac{1}{n}\sum_i X_{Gi} Y_i, \quad \text{and} \\ \widetilde{\beta}_G &= \sum_{g \in G} w_g \widetilde{\beta}_g = \frac{1}{n}\sum_i X_{Gi} \widetilde{Y}_i. \end{aligned}$$

Our permuted linear test statistic is $\widetilde{T}_{G,w} = \widetilde{\beta}_G$, with

$$\operatorname{var}\left(\widetilde{T}_{G,w}\right) = \operatorname{var}\left(\widetilde{\beta}_G\right) = \tfrac{\mu_2}{n-1}\bar{X}_{GG}. \tag{8}$$

For the beta approximation, we need the range of $\widetilde{T}_{G,w}$. Let the sorted Y values be $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(n)}$ and the sorted X_{Gi} values be $X_{G(1)} \leqslant X_{G(2)} \leqslant \ldots \leqslant X_{G(n)}$. Then the range of $\widetilde{T}_{G,w}$ is $[A, B]$, where

$$A = \tfrac{1}{n}\sum_{i=1}^{n} X_{G(i)} Y_{(n+1-i)}, \quad \text{and} \quad B = \tfrac{1}{n}\sum_{i=1}^{n} X_{G(i)} Y_{(i)}.$$

For a $\sigma t_{(\nu)}$ reference distribution we would also need $\mathbb{E}\left(\widetilde{T}_{G,w}^4\right) = \mathbb{E}\left(\widetilde{\beta}_G^4\right)$. We can apply Lemma 2 to the pseudo-gene resulting in

$$\mathbb{E}(\widetilde{\beta}_G^4) = \begin{pmatrix}\mu_2^2 \\ \mu_4\end{pmatrix} A^{\mathsf{T}} B \begin{pmatrix} 3\bar{X}_{GG}^2/n^2 \\ \bar{X}_{GGGG}/n^3 \end{pmatrix}, \tag{9}$$

where $\bar{X}_{GGGG} = \frac{1}{n}\sum_{i=1}^{n} X_{Gi}^4$.

We considered using a $\sigma t_{(\nu)}$ reference distribution for $\widetilde{T}_{G,w}$, taking into account the fourth moment of $\widetilde{T}_{G,w}$ (9). We have often (in fact usually) found that $\mathbb{E}\left(\widetilde{T}_{G,w}^4\right) < 3\mathbb{E}\left(\widetilde{T}_{G,w}^2\right)^2$; that is, lighter tails than the normal. This implies a negative kurtosis for the permutation distribution, and t distributions have positive kurtosis. For this reason we use a beta approximation and not a t approximation.

For the quadratic statistic we have found it useful to replace X_{gi} by $\sqrt{w_g}X_{gi}$ in precomputation. That step is only valid for non-negative w_g, but those are the ones of most interest. Note that mixing positive and negative w_g's would lead to a test statistic where evidence that gene g is non-null could cancel out the evidence of gene h being non-null for $g, h \in G$. Then we use formulas for $\mathbb{E}\left(\widetilde{C}_{G,w}\right)$ and $\operatorname{var}\left(\widetilde{C}_{G,w}\right)$ with all $w_g = w_h = 1$ (5).

Now we consider the computational cost. The cost to compute all of the X_{Gi} is dominated by np multiplications. It then takes n more multiplications to get $\hat{\beta}_G$ and another n to get $\bar{X}_{GGe}$. It costs n multiplications to get μ_2 and μ_4. That step can be done once and can be used for all gene sets. The cost for the Gaussian approximation $\mathcal{N}\left(0, \operatorname{var}(\widetilde{T}_{G,w})\right)$ is dominated by $n(p+2)$ multiplications.

For the beta approximation there is also a cost proportional to $n\log(n)$ in the sorting to compute limits A and B. That adds a cost comparable to a multiple of $\log(n)$ permutations. We judge that the cost of sorting is usually minor for n and p of interest in bioinformatics.

A permutation analysis requires nM multiplications, after computing X_{Gi}, for a total of $n(M+p)$. It is very common for p to be a few tens and M to be many thousands or more. Then we can simplify the costs to $n(M+p) \approx nM$ and $n(2+p) \approx np$. The moment method costs about as much as doing p permutations. When the gene set has tens of genes and the permutation method uses many thousands or even several million permutations, the computational cost is quite large.

The pseudo-gene technique is more expensive for the quadratic statistics. The dominant cost in computing $\widehat{C}_{G,w}$ is still the np multiplications required to compute $\hat{\beta}_g$ for $g \in G$. We can also compute $\mathbb{E}(\widetilde{C}_{G,w})$ in about this amount of work.

The cost of computing $\operatorname{var}(\widetilde{C}_{G,w})$ by a straightforward algorithm is at least np^2, because we need $\bar{X}_{gh}$ and $\bar{X}_{gghh}$ for all $g, h \in G$. Some parts of that computation can be sped up to $O(np)$ by rewriting the expression as described below. One of the terms however does not reduce to $O(np)$. A straightforward implementation costs $O(np^2)$ while an alternative expression costs $O(n^2p)$. The latter is valuable in settings where the gene sets are large compared to the sample size. In the former case, the moment approximation has cost comparable to $O(p^2)$ permutations. If $n < p$ then the latter case is like np permutations, so the quadratic cost is comparable to on the order of $p * \min(n, p)$ permutations.

Recall from Corollary 2 that in an experiment with $n \geqslant 4$ and genes g, h,

$$\operatorname{cov}\left(\widetilde{\beta}_g^2, \widetilde{\beta}_h^2\right) = \begin{pmatrix}\mu_2^2 \\ \mu_4\end{pmatrix}^{\mathsf{T}} A^{\mathsf{T}} B \begin{pmatrix} \bar{X}_{gghh}^*/n^2 \\ \bar{X}_{gghh}/n^3 \end{pmatrix} - \frac{\mu_2^2}{(n-1)^2}\bar{X}_{gg}\bar{X}_{hh},$$

where $\bar{X}_{gghh}^* = \bar{X}_{gg}\bar{X}_{hh} + 2\bar{X}_{gh}^2$ and $A^{\mathsf{T}}B$ is a given 2×2 matrix.

To compute

$$\operatorname{var}\left(\widetilde{C}_{G,w}\right) = \sum_{g \in G}\sum_{h \in G} w_g w_h \operatorname{cov}\left(\widetilde{\beta}_g^2, \widetilde{\beta}_h^2\right)$$

we need μ_2, μ_4 and $A^{\mathsf{T}}B$ which are very inexpensive. We also need

$$S_1 \equiv \sum_{g\in G}\sum_{h\in G} w_g w_h \bar{X}_{gg}\bar{X}_{hh} = \left(\sum_{g\in G} w_g \bar{X}_{gg}\right)^2.$$

By expressing S_1 as a square, we find that it can be computed in $O(np)$ work, not $O(np^2)$ which a naive implementation would provide. We can compute all of the $\bar{X}_{gg}$'s in np multiplications and this is the largest part of the cost. If gene g belongs to many gene sets G we only need to compute $\bar{X}_{gg}$ once and so the cost per additional gene set could be lower.

A similar analysis yields that

$$S_2 \equiv \sum_{g\in G}\sum_{h\in G} w_g w_h \bar{X}_{gghh} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{g\in G} w_g X_{gi}^2\Big)^2$$

is also an $O(np)$ computation. Unfortunately $S_3 \equiv \sum_{g\in G}\sum_{h\in G}\bar{X}_{gh}^2$ does not reduce to an $O(np)$ computation. As written it costs $O(np^2)$. In cases where $p > n$, we can however reduce the cost to $O(n^2p)$ via

$$\begin{aligned} S_3 &= \sum_{g\in G}\sum_{h\in G} w_g w_h \left(\frac{1}{n}\sum_{i=1}^{n} X_{gi}X_{hi}\right)^2 \\ &= \frac{1}{n^2}\sum_{g\in G}\sum_{h\in G} w_g w_h \sum_{i=1}^{n} X_{gi}X_{hj}\sum_{j=1}^{n} X_{gj}X_{hj} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{g\in G} w_g X_{gi}\right)^2. \end{aligned}$$

In terms of these sum quantities,

$$\mathrm{var}(\widetilde{C}_{G,w}) = \begin{pmatrix}\mu_2^2\\ \mu_4\end{pmatrix}^{\mathsf{T}} A^{\mathsf{T}}B\begin{pmatrix}(S_1+2S_3)/n^3\\ S_2/n^3\end{pmatrix} - \frac{\mu_2^2}{(n-1)^2}S_1.$$

Additional file

Additional file 1: Table S1. A table of the moment-based *p*-values for 6,303 gene sets in three genome-wide expression studies.

Abbreviations

GEV: Generalized extreme value; GO: Gene Ontology; GSEA: Gene set enrichment analysis; PD: Parkinson's disease.

Competing interests

JLL is funded by Genentech, Inc. ABO was supported by Genentech, Inc. and by Stanford University while on a sabbatical.

Authors' contributions

JLL and ABO developed the method and wrote the manuscript. JLL implemented the method in the Parkinson's disease data sets. ABO wrote the theoretical sections. Both authors read and approved the final manuscript.

Acknowledgements

We thank Nicholas Lewin-Koh, Joshua Kaminker, Richard Bourgon, Sarah Kummerfeld, Thomas Sandmann, and John Robinson for helpful comments. ABO thanks Robert Gentleman, Jennifer Kesler, and other members of the Bioinformatics and Computational Biology Department at Genentech for their hospitality during his sabbatical there. We also thank two anonymous reviewers for comments that helped us improve our manuscript and method.

Author details

[1]Department of Bioinformatics and Computational Biology, Genentech, Inc., South San Francisco, USA. [2]Currently at GenePeeks, Inc., Cambridge, USA. [3]Department of Statistics, Stanford University, Stanford, USA.

Received: 30 July 2014 Accepted: 10 April 2015
Published online: 28 April 2015

References

1. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267–73.
2. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat. 2007;1:85–106.
3. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci. 2005;102(38):13544–49.
4. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23(8):980–7.
5. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007;23(3):306–13.
6. Lehmann EL, Romano JP. Testing statistical hypotheses. New York: Springer; 2005.
7. Eden T, Yates F. On the validity of Fisher's *z*-test when applied to an actual sample of non-normal values. J Agric Sci. 1933;23:6–7.
8. David HA. The beginnings of randomization tests. Am Statistician. 2008;62(1):70–2.
9. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. Roast: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26(17):2176–82.
10. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics. 2009;10:1–20.
11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
12. Smyth G. Limma: linear models for microarray data In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p. 397–420.
13. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–25.
14. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics. 2001;17(6):509–19.
15. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci. 2001;98(9):5116–121.
16. Bhatia R, Davis C. A better bound on the variance. Am Math Mon. 2000;107(4):353–7.
17. Zhou C, Wang HJ, Wang YM. Efficient moments-based permutation tests. Adv Neural Inf Process Syst. 2009;22:2277.
18. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:80.
19. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RKB, Graeber MB. Whole genome expression profiling of the medial and lateral substantia nigra in parkinsons disease. Neurogenetics. 2006;7(1):1–11.
20. Zhang Y, James M, Middleton FA, Davis RL. Transcriptional analysis of multiple brain regions in parkinsons disease supports the involvement of

specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet. 2005;137B(1):5–16.

21. Scherzer CR, AC ACE, Morse LJ, Liao Z, Locascio JJ, Fefer D, et al. Molecular markers of early Parkinson's disease based on gene expression in blood. Proc Natl Acad Sci. 2007;104(3):955–60.
22. Abou-Sleiman P, Muqit M, Wood N. Expanding insights of mitochondrial dysfunction in parkinsons disease. Nat Rev Neurosci. 2006;7:207–19.
23. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1): 25–9.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological). 1995;57(1):289–300.
26. Fox S, Chuang M, Brotchie J. Serotonin and parkinsons disease: on movement, mood, and madness. Mov Disord. 2009;24(9):1255–66.
27. Berthier J, Jimenez-Sainz A, Pulido R. Pink1 regulates histone h3 trimethylation and gene expression by interaction with the polycomb protein eed/wait1. Proc Natl Acad Sci USA. 2013;110(36):14729–34.
28. Stone D, Reynolds A, Mosely R, Gendelman H. Innate and adaptive immunity for the pathobiology of parkinsons disease. Antioxid Redox Signal. 2009;11(9):2151–66.
29. Good PI. Permutation, parametric, and bootstrap tests of hypotheses. New York: Springer; 2004.
30. Bahadur RR, Savage LJ. The nonexistence of certain statistical procedures in nonparametric problems. Ann Math Stat. 1956;27(4):1115–22.
31. Phipson B, Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9(1):.
32. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate p-values. Bioinformatics. 2009;25(12):161–8.
33. Owen AB. Variance of the number of false discoveries. J R Stat Soc Ser B. 2005;67(3):411–26.
34. Gleich DF, Owen AB. Moment-based estimation of stochastic Kronecker graph parameters. Internet Math. 2012;8(3):232–56.
35. Wedderburn RWM. Random rotations and multivariate normal simulation. Tech Rep. Rothamsted Experimental Station. 1975.
36. Langsrud O. Rotation tests. Stat Comput. 2005;15:53–60.
37. Anderson TW, Olkin I, Underhill LG. Generation of random orthogonal matrices. SIAM J Sci Stat Comput. 1987;8(4):625–9.