Statistical power of gene-set enrichment analysis is a function of gene set correlation structure

DAVID M. SWANSON*

Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, NO 0407

david.swanson@medisin.uio.no

SUMMARY

We develop an analytic statistical framework for examining a variety of gene-set enrichment analysis tests. Within this framework, we describe why statistical power for both self-contained and competitive gene set tests is a function of the correlation structure of co-expressed genes, and why this characteristic is undesireable for gene-set analyses. We additionally describe why past gene-set tests have suffered from inflated type 1 error, and how permutation-based methods have sought to address the issue with some success in the case of self-contained tests and with less success in the case of competitive tests. While the context of this investigation is microarray analysis, with particular focus on leading tests CAMERA, ROAST, SAFE, and GAGE, the observations are also relevant to recently proposed RNAseq gene-set tests, including MAST.

The variable statistical power we describe as a function of gene correlation structure has not been studied. While type 1 error inflation has been well-studied and described previously for both self-contained and competitive tests, it has less often been done in an analytical framework and so it is useful to make assumptions explicit and examine parametric distributions.

We propose three alternative tests, one of which replicates the properties of permutation-based self-contained tests but obviates the need for even recently proposed, rotation-based approximations to burdensome permutations in favor of closed-form densities. The two other tests we propose have the unique property that their statistical power is not a function of co-expression correlation in the gene-set and therefore may be the preferred methodology. We provide simulation support for these proposed methods, compare their results to leading gene-set tests, and apply them to an already-published study of smoking exposure on pregnant women. We call the suite of three proposed test "JAGST" – Just Another Gene-Set Test – and make the methods accessible via an R package of the same name.

Key words: statistical power;GSEA;gene-set testing;pathway analysis;differential expression

1. Introduction

Gene set enrichment analysis (GSEA) or gene-set testing is a class of methods whose goal is to asses the joint enrichment of a biologically interpretable set of genes in a microarray or RNAseq experiment (Subramanian *and others*, 2005; Goeman *and others*, 2004; Kim and Volsky, 2005; Irizarry *and others*, 2009; Luo *and others*, 2009). A variety of methods have been proposed nearly since the advent of microarrays, all with unique advantages, and there have been more recent modifications of these methods as RNAseq becomes a more common platform (Finak *and others*, 2015). There have been two primary categories of gene-set tests, "competitive" and "self-contained" (Wu and Smyth, 2012; Goeman and Bühlmann, 2007; Zhou *and others*, 2013; Rahmatallah *and others*, 2012). The former tries to answer whether genes in a set are more differentially expressed (DE) than some background level of DE on the array. The latter tries to answer whether the gene-set is more DE than one were to expect under the null of no association between transcript abundance and condition. The null distributions for these two kinds

of hypothesis tests have often been calculated via permutation, either at the gene-level for the competitive test, or sample level, for the self-contained test. Manual permutation is used in an attempt to attain the nominal type 1 error rate (Naeem *and others*, 2012; Barry *and others*, 2005) or sometimes quicker methods that yield permutation-like results (Wu *and others*, 2010; Zhou *and others*, 2013). Rejecting a competitive gene-set test test is often a higher bar or more difficult burden of proof, and these tests have been more commonly used in gene-set testing. A rejection of a competitive gene-set test can be more biologically meaningful than a self-contained test.

The use of some GSEA methods has been far more common than others; The seminal paper of Subramanian *and others* (2005) proposed a means of testing a set of genes that leverages the Kolmogorov-Smirnov hypothesis test for similar distributions. It is still often used today and even implemented via a web portal. In recent years, GSEA tests have been proposed that claim greater power and control of type 1 error rates (Rahmatallah *and others*, 2012; Mesirov *and others*, 2016; Wu and Smyth, 2012; Ritchie *and others*, 2015). Error rates have been an issue in some GSEA tests due to correlated test statistics comprising the gene-set test (Goeman and Bühlmann, 2007; Wu and Smyth, 2012; Gatti *and others*, 2010; Ritchie *and others*, 2015; Mesirov *and others*, 2016).

In this paper, we describe within an analytic statistical framework why past gene-set tests have suffered from inflated type 1 error, and how permutation-based methods have sought to address the issue in different ways. Goeman and Bühlmann (2007) made a similar and rigorous description of gene-set tests, though in the context primarily of 2×2 tables, and prior to the development of some of the tests we analyze. More importantly, we additionally describe in the same framework why past and more recently proposed gene-set tests suffer from having variable power as a function of gene set correlation structure.

We show that the location of causal transcripts within the co-expression correlation structure

is critical in determining statistical power of the test in addition to the effect sizes of them, and how the issue is relevant to both competitive and self-contained gene-set tests. Briefly, correlated blocks of genes manifest themselves in disproportionate degree in the tails of distributions under both the null and alternative hypotheses. Since we generally perform statistical tests in the tails of distributions, hypothesis tests are disproportionately influenced by this correlation. When causal transcripts are found in these correlated blocks, power for their detection is likewise overly represented relative to causal transcripts found in less correlated. This observation has been made in the context of region-based SNP testing, though not in the context of gene-set testing (Swanson *and others*, 2013). Despite the observation being made in the context of SNP-set testing, it is pertinent to mention since tests continue to be proposed that suffer from variable power under different correlation structures (Bakshi *and others*, 2016).

We leverage our statistical analysis of gene-set tests by proposing three tests, the suite of which we call "JAGST" (Just Another Gene Set Test). The first test is self-contained and replicates results from a leading self-contained test ROAST, but does so using a closed form mixture density rather than the rotation and iteration-based method of ROAST. Though ROAST is an elegant, general, and computationally cheap algorithm for various set-based hypothesis tests, two of its more important special cases (mixed and directed set-based tests) fall within our statistical framework so that relevant null tail probabilities can be written down and calculated.

The second two tests of JAGST are power-invariant self-contained and competitive tests. The proposed JAGST power-invariant self-contained test simply uses correlation information of DE test statistics to implement a standard chi-square test. The proposed competitive test is more complex and computationally intensive, using a sophisticated and recently proposed high-dimensional penalized regression method for obtaining z-statistics from a variable selection procedure (Zou *and others*, 2007; Lockhart *and others*, 2014; Taylor and Tibshirani, 2017; Tibshirani *and others*, 2016). We use these z-statistics to create an appropriate mixture distribution of non-central

chi-squares, which is the null distribution for our competitive test.

We perform simulations with leading and commonly used competitive and self-contained tests, including ROAST, CAMERA, SAFE, and GAGE (Wu and Smyth, 2012; Wu *and others*, 2010; Barry *and others*, 2005; Luo *and others*, 2009) to demonstrate the degree to which statistical power is effected by correlation structure of the gene-set and how our proposed tests do not have this same property. ROAST and SAFE are self-contained tests, while CAMERA and GAGE are competitive tests. We perform a data analysis on an already published study of smoking exposure in pregnant women (Votavova *and others*, 2011). While expression studies are the context of our analysis, our observations are also relevant to RNAseq. The popular MAST method for RNAseq is based on CAMERA, and many of the same statistical principles will apply in the RNAseq context (Finak *and others*, 2015). Additionally, expression analyses continue to be a valuable and well-used platform for diagnosis and prediction (Tachibana, 2015; Zhao *and others*, 2014; Byron *and others*, 2016; Zhang *and others*, 2015; Xu *and others*, 2016).

2. Methods

Consider matrix Y ($n \times m$) of normalized transcript measures and matrix X ($n \times p$) of conditions, where Y has $m \times m$ correlation matrix Σ, whose i, j^{th} element is $\rho_{i.j}$. Assume for now that $p = 1$ for simplicity. It is common in GSEA to first perform a standard microarray analysis, and then test gene-set enrichment on the summary statistics (such as transcript-level test statistics).

If we fit a regression model of the i^{th} column of Y (Y_i) on X, $i \in 1 \dots m$, we extract some test statistic, call it T_i (oftentimes a t-statistic for linear regression if the outcome is continuous), associated with this model, and likewise T_j for the model of Y_j on X. The test statistic from a linear regression model is

$$T_i \equiv \frac{Y_i^T \cdot X}{\sqrt{\hat{\sigma}^2 X^T \cdot X}}$$

if we have centered X and Y, which follows a t-distribution on n-2 df under the null hypothesis.

Likewise, define

$$T_j \equiv \frac{Y_j^T \cdot X}{\sqrt{\hat{\sigma}^2 X^T \cdot X}}.$$

Since $Cor(\cdot, \cdot)$ is invariant to scaling its arguments and assuming centered, normalized Y_i's, Y_j's, and X's,

$$Cor(T_i, T_j) = E(T_i \cdot T_j) = E(Y_i \cdot X \cdot Y_j \cdot X) = E(Y_i \cdot Y_j) \cdot E(X^2) = \rho_{ij}$$

where the second to last equality holds because we calculate under the null of no association with X. So we see that $Cor(T_i, T_j)$ is the (i, j) entry of Σ, the correlation matrix of Y. It follows that the vector T composed of entries T_k, $k = 1, \ldots, m$ also has $Cor(T) = \Sigma$.

Now consider the regression model under the alternative hypothesis of an association between a condition and transcript abundance. In particular, suppose there exists a causal association between X and Y_j such that the expectation of the test statistics T_j corresponding to regression model $Y_j \sim X$ is $E[T_j] = \mu_j$. For example, power under this association for an α-level test would be calculated with

$$F_{t_{n-2}}(\mu_j + q) + (1 - F_{t_{n-2}}(\mu_j - q))$$

where $F_{t_{n-2}}(\cdot)$ is the cdf of a t-distribution on n-2 df and q is the $\alpha/2$ quantile of that distribution.

Assume there exists no causal association between X and Y_i, and consider test statistic T_i. Because $Cor(T_i, T_j) = \rho_{i,j}$, $E(T_i) = \rho_{i,j} \cdot \mu_j$; we have power to detect an association between X and Y_i by virtue of Y_i's correlation with Y_j. though less so because $|\rho_{i,j}| \leqslant 1$. We will return to this point later in the discussion of statistical power.

Suppose that we have sufficient sample size such that the t-statistics resulting from an expression analysis can be approximated by the standard normal distribution. Since the test statistic for many gene-set tests relies on or is highly correlated with a sum or sum of squares of probe-level summary statistics composing the gene-set (e.g, Wu *and others* (2010); Kim and Volsky

(2005); Wu and Smyth (2012); Ritchie *and others* (2015); Luo *and others* (2009); Barry *and others* (2005)), consider the following property about the multivariate normal distribution.

$$\sum_{i\in\Gamma} T_i = J_G^T \cdot T \sim N(J_G^T \cdot \mu, J_G^T \, \Sigma \, J_G) \tag{2.1}$$

for T a length g vector that follows $MVN(\mu, \Sigma)$, Γ a size g set of indices for the gene-set, and J_G an indicator vector whose 1 entries correspond to the indices of Γ. So the mean of $\sum_{i\in\Gamma} T_i$ is the sum of the component means corresponding to the 1 entries in J_G, and variance the sum of the entries of Σ that remain after multiplication by J_G. Sometimes this test statistic in the gene-set testing context is called the directed or directional test since its value is sensitive to the sign of T_i.

Additionally, if we square the T_i's, we have

$$\sum_{i\in\Gamma} T_i^2 = \sum_{i\in\Gamma} \lambda_i \cdot \chi_1^2(\delta_i^2) \text{ where } \delta_i = (\mu_{\mathbf{J}}{}^T \cdot \nu_{\mathbf{i}})/\sqrt{\lambda_i}, \tag{2.2}$$

$\mu_J \equiv J_G \cdot \mu$, and λ_i and $\nu_{\mathbf{i}}$ are the eigenvalues and eigenvectors of $J_G^T \, \Sigma \, J_G$ (Imhof, 1961; Press, 1966; Harville, 1971). Sometimes this test statistic in the gene-set testing context is called undirected, non-directional, or mixed as its power is invariant to the sign of T_i.

3. Relationship to permutation-based tests

3.1 *Considering preservation of Type 1 error using permutations*

It is sometimes thought that permutation is a sure, if computationally burdensome, way of preserving type 1 error when the null distribution of a test statistic is unknown. We show why this is not the case, at least with competitive gene-set tests. The result has been noted frequently elsewhere (Wu and Smyth, 2012; Ritchie *and others*, 2015; Zhou *and others*, 2013), but we try to describe it in a statistical framework so that evidence of the phenomenon is not primar-

ily simulation-based. First, however, we consider permutation-based self-contained tests and the parametric distributions approximated by them.

3.1.1 *Self-contained tests type 1 error* In light of the results in Section 2, we can consider the null distributions generated when samples ("self-contained" gene-set tests) or genes ("competitive" gene-set tests) are permuted to generate a null distribution for the test statistics discussed. In permuting samples, any possible association between gene and outcome is broken so that the expectation of T, the test statistic vector described above, is a length g vector with expectation $\mathbf{0}$, rather than μ. Since the permutations do not permute the particular gene set under consideration, their own correlation structure is preserved. Thus, one can conclude that a permutation-generated null distribution is a sample from

$$T_{null,self} \sim MVN(0, \Sigma_G), \text{ where } \Sigma_G \equiv J_G^T \Sigma J_G.$$

With the joint distribution of $T_{null,self}$ in mind, we can consider the sum of squares of its elements. Since the expectation of $T_{null,self}$ is zero and using notation conventions from Section 2, $\delta_i = 0$ because ν_i will be an inner product with a $\mathbf{0}$ vector. Differences in the null distribution for $\sum_{i\in\Gamma} T_i^2$ are therefore driven by $\{\lambda_i\}$ (the set of eigenvalues of Σ_G), and the null distribution can be very different depending on the distribution of the $\{\lambda_i\}$'s. If there is high within-set correlation, greater variation in λ_i will lead to a heavier-tailed null, whereas little within-set correlation will correspond to relatively similar λ_i's and a thinner tail.

No within gene-set correlation implies independent T_i's, so that

$$\sum_{i=1}^{n} T_i^2 \sim \chi_g^2.$$

It is because independent component test statistics have been assumed at times in the past that a χ_g^2 null distribution is used when in fact the heavier tailed $\sum_{i=1}^{n} \lambda_i \cdot X_1^2$ was the true null distribution (Irizarry *and others*, 2009). It is in this distinct way that even self-contained tests

have occasionally had inflated type 1 error, though this has been recognized and corrected in different articles (Gatti *and others*, 2010; Wu *and others*, 2010; Zhou *and others*, 2013; Mesirov *and others*, 2016).

3.1.2 *Competitive tests type 1 error* When permuting across gene sets for a competitive gene set test, probe-wise associations with the outcome are maintained, but the set over which we aggregate changes as a function of the genes randomly chosen to compose our permutation-generated set. Thus, for permutation k of the null generation, the distribution from which we sample is

$$T_{null,comp,k} \sim MVN(\mu_{G_k}, \Sigma_{G_k}),$$

where G_k is a set of genes of size g randomly chosen, μ_{G_k} is the expectation vector of test statistics corresponding to those particular genes, and Σ_{G_k} is a submatrix of Σ corresponding to the set G_k. Since G_k changes with each iteration, it is evident that $\{T_{null,comp,1}\ , \ldots,\ T_{null,comp,q}\}$, where q is the size of the permutation-generated null, are not identically distributed, but a mixture of distributions.

Crucially, and likely the reason type 1 error has been difficult to maintain in competitive gene-set tests, the off-diagonal entries of Σ_{G_k} (the correlation matrix of permutation k) will often be systematically smaller than the off-diagonal entries of Σ_G because genes within biologically meaningful sets will often be co-expressed to a greater degree than randomly chosen groups of genes. We may therefore expect that the permutation-generated null distribution will not be as thick tailed as the test statistic whose null it is trying to approximate.

3.2 *Statistical power when using permutations*

Because expression analyses often rely on univariate regression test statistics, confounding structures are not detected and power to detect some non-causal transcripts can be above α because

of their association with the causal gene. The underlying causal structure of transcript on phenotype and confounding by co-expressed genes and their effects on statistical power has been less discussed on a methodological level. When variable power has been discussed, it has been done so from the perspective of underlying biological or platform phenomena, such as in Oshlack and Wakefield (2009) with respect to transcript length bias. Here we describe why statistical power for both self-contained and competitive gene-set tests can vary due to the methodological set-up of the tests themselves.

3.2.1 *Illustrative example* We first consider an example to explain variable power more concretely. For simplicity, assume a gene-set consists of 3 genes, one of which has a causal association with the outcome, and the other two do not. Call the test statistic associated with the causal gene T_1 with expectation μ_1, and those of the non-causal genes T_2 and T_3 (both with expectations 0). First suppose that T_2 and T_3 have correlation ρ, but both are independent of T_1 (the causal gene). The distribution of their sum is

$$T_1 + T_2 + T_3 \sim N(\mu_1, 3 + 2\rho),$$

while the non-directional test statistic is

$$T_1^2 + T_2^2 + T_3^2 \sim \sum_{i=1}^{3} \lambda_i \cdot \chi_1^2(\delta_i^2) \text{ with. } \delta_i = (\mu_1 \;\; 0 \;\; 0) \cdot \nu_{\mathbf{i}}$$

and λ_i and $\nu_{\mathbf{i}}$ are the eigenvalues and eigenvectors of Σ.

Consider now the scenario where we "shift" the correlation structure so that T_1 and T_2 have correlation ρ, and T_3 is independent of them both. though the same causal associations remain. So $E(T_1) = \mu_1$, and in this scenario $E(T_2) = \mu_1\rho$ because of T_2's correlation with T_1. The expectation of T_3 is still 0. This time, the distribution of their sum is

$$T_1 + T_2 + T_3 \sim N(\mu_1(1 + \rho), 3 + 2\rho)$$

and the non-directional test statistic is

$$T_1^2 + T_2^2 + T_3^2 \sim \sum_{i=1}^{3} \lambda_i \cdot \chi_1^2(\delta_i^{*2}) \text{ where } \delta_i^* = (\mu_1 \;\; \mu_1\rho \;\; 0) \cdot \nu_{\mathbf{i}}.$$

It will always be the case that $\delta_i^{*2} \geqslant \delta_i^2 \;\; \forall \;\; i$, that is, in the second scenario the non-centrality parameters will tend to be bigger than those in the first scenario. We know this because $(\mu_1 \;\; \mu_1\rho \;\; 0)$ is greater than $(\mu_1 \;\; 0 \;\; 0)$ element-wise. Additionally, there will exist at least one δ_i strictly greater in the second scenario than the first scenario because for at least one i, ν_i has a non-zero entry in the second element since eigenvectors span the space defined by the columns of Σ. Because χ^2 distributions are stochastically strictly increasing in their non-centrality parameters, we will always be better powered to detect the gene-set in the second scenario than in the first.

From this example, we conclude that we are more likely to reject the null for gene set enrichment test when causal transcripts are co-expressed with non-causal ones.

4. Proposed hypothesis tests: JAGST

4.1 *Proposed hypothesis test 1: an analytic approach to self-contained test ROAST*

First we introduce a test that is nearly numerically equivalent to important special cases of a leading self-contained gene-set test ROAST, but whose implementation relies on closed form formulae rather than the rotation-based methodology of ROAST, itself an elegant way of approximating the results of permutation-based procedures.

In Section 2 we gave formulae in equations (1) and (2) for the sum of test statistics and sum of squares of test statistics, respectively. These two distributions correspond to two important special cases of ROAST – its directional and mixed gene-set tests. We therefore do not need to rely on the rotation of residual methodology, which though statistically elegant has p-value granularity dependent on the number of specified iterations. We can instead calculate p-values

using tail probabilities of the normal distribution in the case of the directional test and of a mixture of scaled χ^2 distributions in the case of the mixed test.

We see in the supplementary material the correlation of 0.99 (0.96) between the ROAST directional test (mixed test) and analytic calculation of tail probabilities using these densities, both under the null hypothesis. That correlation falls slightly under the alternative hypothesis for reasons explained in the figures, but remains above 0.88. If the variable power issue is not a concern for the analyst, for example for reasons given in the Discussion section, and p-values from the self-contained test ROAST methodology are preferred, one need only use these formulae given. Since inversion of the χ^2 mixture distribution can be difficult, sampling from the mixture distribution is also adequate and likely still more computationally inexpensive than ROAST.

4.2 *Proposed hypothesis test 2: power-invariant competitive gene-set test*

We propose the JAGST competitive test whose statistical power is invariant to the correlation structure in which causal effects are found. We do so by making two key changes to currently used gene-set testing methods. First, the test statistic for the competitive test is calculated using

$$T_G^T \, \Sigma_G^{-1} \, T_G,$$

rather than a straightforward sum or sum of squares of test statistics. Using the inverse covariance matrix of the test statistics is central to achieving the desired correlation structure power invariance. Secondly, the null distribution is calculated by taking random subsets of genes of size g and calculating the same test statistic. Unlike other competitive tests that calculate their null distribution with random subsets of genes, using the inverse correlation matrix makes each realization of our null distribution a function of only the underlying effects of each subset thereby controlling for correlation structure.

While this is the essential idea of the JAGST competitive test, in practice we generate the proposed null distribution in a different and computationally easier way. Especially for large gene-

sets, some of which approach a few hundred genes, calculating the null distribution in this brute force way is too cumbersome if not numerically prohibitive depending on whether the correlation matrices are approximately singular.

We therefore propose instead to:

1. Perform variable selection on the random gene subsets of size g,
2. choose the model with the smallest AIC (or another model fitness criterion that could be shown to control type 1 error and be power-invariant), then
3. use a novel and sophisticated method to calculate the z-statistics of the selected variables (Taylor and Tibshirani, 2017; Lockhart *and others*, 2014; Tibshirani *and others*, 2016), in order to
4. take the sum of squares of these test statistics, calling them $\delta_{PI} \equiv \sum \delta_{i,PI}^2$ ("PI" for "power-invariant"), and
5. sample from $\chi_g^2(\delta_{PI})$, which is the null distribution of the above test statistic as demonstrated in simulation (Figure 4).

We iterate until desired granularity in the null distribution is achieved.

The L1-penalized regression hypothesis testing procedure of Taylor and Tibshirani (2017) proves central to our algorithm. It provides a means of obtaining z-statistics on selected variables in a penalized regression framework and until recently was not possible. These z-statistics allow us to determine the distribution under the alternative that our DE test statistics arise from.

While variable selection is not without computation cost, since the procedure stops once a certain AIC is achieved, we will generally avoid the large $O(g^3)$ cost of matrix inversion and numeric instability for large gene sets by a significant margin. This is especially true if it's assumed that in general gene sets are composed of many transcripts working together in pathways, though reveal relatively sparse models when regularization is applied.

Iteration over different sets of size g will yield a mixture of non-central χ^2 distributions. Formulae and approximations have been proposed for mixtures of χ^2 distributions (Press, 1966; Liu *and others*, 2009), which could be used if the same null distribution were relevant to many different tests, or sampling from distributions were burdensome. For our purposes, we eschew these formulae to prefer sampling from the mixture distribution in R (R Core Team, 2017).

While other model fitness criteria could be used, we propose AIC at least in part because it chooses less parsimonious models than, for example, BIC. Since models are less parsimonious, the procedure we propose is less likely to be anti-conservative since δ_{PI}, the sum of squares of non-centrality parameters, will tend to be larger. Indeed, simulation suggests that type 1 error is controlled, without being overly conservative. If other simulation scenarios suggested otherwise, different model fitness criteria could be used.

4.3 *Proposed hypothesis test 3: power-invariant self-contained gene-set test*

For the JAGST self-contained test, we again use the test statistic

$$T_G^T \, \Sigma_G^{-1} \, T_G.$$

In this case, the null hypothesis assumes no association between transcript and outcome so the null distribution for this test statistic is simply χ^2_g. The quantiles for this null distribution will often be a much lower threshold for statistical significance as compared to the proposed competitive gene set test.

5. Results

We provide simulation and data analysis results for the different JAGST tests and compare them with CAMERA (with and without ranks), GAGE (with and without ranks), ROAST, and SAFE. We abbreviate our simulation analysis of type 1 error since it has been described well and studied

in detail elsewhere (Barry *and others*, 2005, 2008; Zhou *and others*, 2013; Wu and Smyth, 2012) and focus on power as a function of correlation structure under different generating models. We perform simulation and data analysis in the R language (R Core Team, 2017).

5.1 *Simulation*

We generated 280 samples, each with 40 transcripts or features, the 40 composed of a correlated region of size 20 with correlation 0.8 and an uncorrelated region of size 20. The correlated and uncorrelated blocks also compose our two hypothetical gene sets. We then generated two different binary outcomes consistent with a logistic regression model, where the probability of event in one case was a function of two transcripts in the correlated block and in the other case a function of two uncorrelated transcripts in the uncorrelated block. The effect sizes of the two transcripts were equal and were also the same in either scenario (i.e., whether the causal transcripts were in the correlated or uncorrelated blocks).

We varied the effect size of the two transcripts from a log odds ratio of zero to 3 in the correlated region and performed eight different gene set tests on the 20 features composing the correlated block: CAMERA (with and without ranks), GAGE (with and without ranks), ROAST, SAFE, and the two JAGST power-invariant tests were propose, one self-contained and the other competitive.

We then varied the effect sizes of the two transcripts in the uncorrelated block over the same values and performed the same eight gene tests, this time on the 20 features composing the uncorrelated block. The two power curves in each of Figures 1-6 correspond to the respective test applied to the 20 correlated or uncorrelated features and the associated outcome.

In the cases of CAMERA, ROAST, SAFE, and GAGE, at all effect sizes greater than zero, there was more power to detect the correlated block gene set, even though effect sizes were the same as compared to the uncorrelated block gene set (Figures 1-6). Indeed, the power curves

diverged and had different slopes on the -log p-value scale so that there was a still bigger power differential at increasing effect sizes. The divergence is particularly striking with GAGE, whose simulation was repeated to confirm the result. In the cases of Figures 1 and 2, we see a ceiling on the -log p-value, essentially stopping the divergence of power curves. This occurs because the methods use empirical p-values, and we set the number of iterations to 1000. The divergence would continue if the number of iterations were increased. There is a similar ceiling on the rank-based tests found in Figures 4 and 6.

Our proposed gene set tests on the other hand yielded power curves that were much closer at all effect sizes of the two underlying causal transcripts (Figures 7 and 8). While there are small power differences with at least the proposed competitive JAGST test, the power curves were parallel, indicating that the power difference in curves does not increase for larger effect sizes.

Lastly we performed gene-set tests under the null hypothesis and weak alternative to compare results from ROAST's directional tests with tail probabilities justified in Section 2. The alternative hypothesis posited two causal transcripts among a gene-set size of 20 correlated transcripts. Rather than inverting the distribution, we estimated this tail probability using the appropriate mixture distribution of central, scaled χ^2's. Figures 9 and 10 show the high degree of correspondence between the two ways of calculating the test, with a correlation of 0.99. Figures found in the Supplementary Material show the correspondence between the mixed, non-directional tests of ROAST and the analytic solution described in Section 2. Again there is a high degree of correspondence between the two tests under the null and alternative hypotheses, both with correlations above 0.9.

5.2 *Data analysis*

We analyze data from an already published study of maternal and fetal transcription variation due to smoking exposure (Votavova *and others*, 2011). The study analyzed maternal peripheral,

placenta, and neonatal cord blood on 20 pregnant women with smoking exposure and 50 without significant smoking exposure. Women were assayed using Illumina expression beadchip v3 and yielded 24,526 transcripts. We accessed the data via its Gene Expression Omnibus (GEO) accession number GDS3929 (Edgar *and others*, 2002; Barrett *and others*, 2013).

In keeping with the analysis protocol of Votavova *and others* (2011), we filtered transcripts of each cell type to approximately the third at detectable expression level. In doing so, the most differentially expressed genes according to our analyses were consistent with those of Votavova *and others* (2011). For our gene-set analyses, we only used this approximate third most differentially expressed in each category.

We analyzed 7 different gene-sets, motivated by ones commonly used in gene-set enrichment studies, oncogenic pathways, and ones more specific to the context of (Nevins and Potti, 2007; Zhang *and others*, 2010; Bild *and others*, 2006; Votavova *and others*, 2011) (e.g., placenta development). We used 8 different gene-set test methodologies: CAMERA (with and without ranks), ROAST, SAFE, and GAGE (with and without ranks). ROAST and SAFE are self-contained tests, while CAMERA and GAGE are competitive tests. The last two used are our proposed power-invariant competitive and self-contained tests. P-values for the 8 tests by gene-set and data source (maternal peripheral blood, neonatal cord blood, or placenta) are included in Tables 1, 2, and 3, respectively.

6. Discussion

Much has been learned from gene-set testing since the early 2000's when such methods were developed. The idea that because biological processes are complex systems whose elements are not acting in isolation is powerful and has rightly been leveraged in many gene-set testing methodologies. Most of these methods are uniquely suited to certain situations, implicitly or explicitly making different assumptions of the underlying data and biological process.

Additionally, most proposed gene-set tests make much intuitive sense, for example summing test statistics from differential expression analyses and considering correlation structure in those test statistics to varying degrees and in different ways. For example, CAMERA accounts for that correlation structure via a variance inflation factor (VIF), while ROAST does so via generating a null distribution with rotation of residuals. As methodology has become more sophisticated over time, many tests have been proposed as a response to deficiencies identified in existing methods. For example, recognition that without accounting for correlated test statistics some even self-contained gene-set tests had grossly inflated type 1 error rates led to many critical methodological developments (Zhou *and others*, 2013; Wu *and others*, 2010; Mesirov *and others*, 2016). And to the extent that permutation has been proposed as a means of controlling type 1 error in some cases Barry *and others* (2005), other tests have been proposed that avoid the computational demands of permutation (Wu *and others*, 2010; Zhou *and others*, 2013).

We described many gene-set tests within an explicit statistical framework in this paper and showed why some could be subject to type 1 error inflation. While that observation has been made previously, its best description was done in the context of 2 x 2 tables (Goeman and Bühlmann, 2007), and we describe why it can be the case with a linear combination of differential expression statistics.

More importantly in this paper, we observed and explained why all gene-set tests analyzed have variable power as a function of correlation structure. We show that if gene drivers of phenotype are found in uncorrelated regions of a gene-set, there is much less power to detect the gene-set than if the drivers are in a correlated region of the gene-set. The observation is particularly relevant because co-expression of genes on expression arrays is ubiquitous and variable; indeed, the very idea of gene-set testing is that groups of genes working in concert are expressed and can be tested together.

The gene set tests applied in our data analysis of smoking exposure on pregnant women did

not yield any significant results adjusted for multiple testing. Our proposed self-contained and competitive tests were less significant in the cases of nearly every gene set. The proposed JAGST competitive gene set test is less significant than the JAGST self-contained test, which is consistent with the idea that competitive tests are a more significant burden of proof than self-contained tests. While we were able to replicate the DE results of Votavova *and others* (2011), we were not able to confirm findings of significant gene-sets. Since Votavova *and others* (2011) relied on the DAVID database for enrichment analyses, we cannot expect our results to coincide with those of the previous study (Huang *and others*, 2008, 2009) .

The self-contained tests used in our data analysis, ROAST and SAFE, were not generally more significant than the competitive tests, Camera and Gage. Since tests are a function of correlation structure, and in the cases of the ranked tests, non-parametric, we would not necessarily expect an obvious distinction between the two kinds of tests even though their null hypotheses are different.

While we have pointed out power variability as a function of correlation structure in gene-set tests, it is important to point out that the relevance of use of the JAGST methodology will depend on application. When the correlation structure and location of transcripts driving phenotype relative to that structure are similar to those used in our simulation scenario, then there could be significant under-detection of certain gene sets depending on the genetic architecture of the pathway. On the other hand, there are also scenarios, for example when there is only a single correlated block in the gene set, when power variation is likely less of a concern and any of the tests analyzed in this paper are suitable methods. In principle, however, it seems best to use a test that is statistically sound without such implicit assumptions about correlation structure, in which case the JAGST method is preferred.

It is unclear why exactly there is a small divergence in power curves for the JAGST competitive test in the simulation scenarios compared (see Figure 7). Since our method relies on test statistics from an L1-penalized regression model, it is likely the relatively small variation is due to very

different correlation structures from the simulation in which the transcripts with non-zero effect sizes are found. These correlation structures may affect AIC, the model fitness criterion being used to choose the tuning parameter, which then affects magnitude of the test statistics. It is important, though, that that small difference we see in our test is much smaller than that observed in other tests and that additionally the two power curves are parallel on the -log p-value scale. In contrast, we see that divergence in power curves increases with effect size for the other tests examined. For the JAGST self-contained test, we observe nearly identical power curves indicating nearly perfect power invariance.

7. Software and Data

An R package for implementing JAGST is available on GitHub under "sojourningNorth/JAGST" and can be installed in the standard way using directions found at the repository. Additionally, code used for simulation and analysis found in the manuscript is available at "sojourningNorth/JAGST-analysis". Data used for analysis is available through NCBI's GEO database and has accession number GDS3929 (Edgar *and others*, 2002; Barrett *and others*, 2013).

8. Supplementary Material

Additional figures showing correspondences between the directional and mixed tests of ROAST and our calculation of these tests is in the online Supplementary Material and can be found at `http://biostatistics.oxfordjournals.org`.

Acknowledgements

The author wishes to thank Prof. Arnoldo Frigessi for helpful comments in the preparation of the manuscript.

References

Bakshi, Andrew, Zhu, Zhihong, Vinkhuyzen, Anna AE, Hill, W David, McRae, Allan F, Visscher, Peter M and Yang, Jian. (2016). Fast set-based association analysis using summary data from gwas identifies novel gene loci for human complex traits. *Scientific reports* **6**, 32894.

Barrett, Tanya, Wilhite, Stephen E., Ledoux, Pierre, Evangelista, Carlos, Kim, Irene F., Tomashevsky, Maxim, Marshall, Kimberly A., Phillippy, Katherine H., Sherman, Patti M., Holko, Michelle, Yefanov, Andrey, Lee, Hyeseung, Zhang, Naigong, Robertson, Cynthia L., Serova, Nadezhda, Davis, Sean *and others.* (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research* **41**(D1), 991–995.

Barry, William T., Nobel, Andrew B. and Wright, Fred A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics* **21**(9), 1943–1949.

Barry, William T., Nobel, Andrew B. and Wright, Fred A. (2008). A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics* **2**(1), 286–315.

Bild, A H, Yao, G, Chang, J T, Wang, Q, Potti, A, Chasse, D, Joshi, M B, Harpole, D, Lancaster, J M, Berchuck, A, Olson Jr., J A, Marks, J R, Dressman, H K, West, M *and others.* (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**(7074), 353–357.

Byron, Sara A., Van Keuren-Jensen, Kendall R., Engelthaler, David M., Carpten, John D. and Craig, David W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics* **17**(5), 257–271.

Edgar, Ron, Domrachev, Michael and Lash, Alex E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**(1), 207–210.

Finak, Greg, McDavid, Andrew, Yajima, Masanao, Deng, Jingyuan, Gersuk, Vivian, Shalek, Alex K., Slichter, Chloe K., Miller, Hannah W., McElrath, M. Juliana, Prlic, Martin, Linsley, Peter S. *and others*. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**(1), 278.

Gatti, Daniel M, Barry, William T, Nobel, Andrew B, Rusyn, Ivan and Wright, Fred A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC genomics* **11**(1), 574.

Goeman, Jelle J. and Bühlmann, Peter. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23**(8), 980–987.

Goeman, Jelle J, Van De Geer, Sara A, De Kort, Floor and Van Houwelingen, Hans C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**(1), 93–99.

Harville, David. (1971). On the Distribution of Linear Combinations of Non-central Chi-Squares. *The Annals of Mathematical Statistics* **42**(2), 809–811.

Huang, Da Wei, Sherman, Brad T and Lempicki, Richard A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**(1), 44–57.

Huang, Da Wei, Sherman, Brad T. and Lempicki, Richard A. (2009). Bioinformatics en-

richment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**(1), 1–13.

Imhof, J P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**(3/4), 419–426.

Irizarry, R. a., Wang, C., Zhou, Y. and Speed, T. P. (2009). Gene Set Enrichment Analysis Made Simple. *Statistical methods for medical research* **18**(6), 565–575.

Kim, Seon-Young and Volsky, David J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* **6**, 144.

Liu, Huan, Tang, Yongqiang and Zhang, Hao Helen. (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**(4), 853–856.

Lockhart, Richard, Taylor, Jonathan, Tibshirani, Ryan J. and Tibshirani, Robert. (2014). A significance test for the lasso. *Annals of Statistics* **42**(2), 413–468.

Luo, Weijun, Friedman, Michael S, Shedden, Kerby, Hankenson, Kurt D and Woolf, Peter J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics* **10**, 161.

Mesirov, Pablo Tamayo, Steinhardt, George, Liberzon, Arthur and P, Jill. (2016). The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research* **25**(1), 472–487.

Naeem, Haroon, Zimmer, Ralf, Tavakkolkhah, Pegah and Küffner, Robert. (2012). Rigorous assessment of gene set enrichment tests. *Bioinformatics* **28**(11), 1480–1486.

Nevins, Joseph R and Potti, Anil. (2007). Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature reviews. Genetics* **8**(8), 601–609.

Oshlack, A and Wakefield, MJ. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology direct* **4**, 14.

Press, Sheldon James. (1966). Linear combinations of non-central chi-square variates. *The Annals of Mathematical Statistics*, 480–487.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahmatallah, Y., Emmert-Streib, F. and Glazko, G. (2012). Gene set analysis for self-contained tests: Complex null and specific alternative hypotheses. *Bioinformatics* **28**(23), 3073–3080.

Ritchie, Matthew E., Phipson, Belinda, Wu, Di, Hu, Yifang, Law, Charity W., Shi, Wei and Smyth, Gordon K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**(7), e47.

Subramanian, Aravind, Tamayo, Pablo, Mootha, Vamsi K, Mukherjee, Sayan, Ebert, Benjamin L, Gillette, Michael A, Paulovich, Amanda, Pomeroy, Scott L, Golub, Todd R, Lander, Eric S *and others*. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550.

Swanson, David, Blacker, Deborah, AlChawa, Taofik, Ludwig, Kerstin, Mangold, Elisabeth and Lange, Christoph. (2013). Properties of permutation-based gene tests and controlling type 1 error using a summary statistic based gene test. *BMC Genetics* **14**(1), 108.

Tachibana, Chris. (2015). Transcriptomics today: Microarrays, RNA-seq, and more. *Science*.

Taylor, Jonathan and Tibshirani, Robert. (2017). Post-selection inference for ???1-penalized likelihood models. *Canadian Journal of Statistics* (2015), 1–26.

Tibshirani, Ryan J., Taylor, Jonathan, Lockhart, Richard and Tibshirani, Robert. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association* **111**(514), 600–620.

Votavova, H., Dostalova Merkerova, M., Fejglova, K., Vasikova, A., Krejcik, Z., Pastorkova, A., Tabashidze, N., Topinka, J., Veleminsky, M., Sram, R. J. *and others*. (2011). Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta* **32**(10), 763–770.

Wu, Di, Lim, Elgene, Vaillant, François, Asselin-Labat, Marie Liesse, Visvader, Jane E. and Smyth, Gordon K. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**(17), 2176–2182.

Wu, Di and Smyth, Gordon K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research* **40**(17), 1–12.

Xu, Joshua, Gong, Binsheng, Wu, Leihong, Thakkar, Shraddha, Hong, Huixiao and Tong, Weida. (2016). Comprehensive assessments of RNA-seq by the SEQC consortium: FDA-led efforts advance precision medicine. *Pharmaceutics* **8**(1).

Zhang, Jiao, Chen, Yan-Hua and Lu, Qun. (2010). Pro-oncogenic and anti-oncogenic pathways: opportunities and challenges of cancer therapy. *Future oncology (London, England)* **6**(4), 587–603.

Zhang, Wenqian, Yu, Ying, Hertwig, Falk, Thierry-Mieg, Jean, Zhang, Wenwei, Thierry-Mieg, Danielle, Wang, Jian, Furlanello, Cesare, Devanarayan, Viswanath, Cheng, Jie, Deng, Youping, Hero, Barbara, Hong, Huixiao, Jia, Meiwen, Li, Li, Lin, Simon M, Nikolsky, Yuri, Oberthuer, André, Qing, Tao, Su, Zhenqiang, Volland, Ruth, Wang, Charles, Wang, May D, Ai, Junmei, Albanese,

Davide, Asgharzadeh, Shahab, Avigad, Smadar, Bao, Wenjun, Bessarabova, Marina, Brilliant, Murray H, Brors, Benedikt, Chierici, Marco, Chu, Tzu-Ming, Zhang, Jibin, Grundy, Richard G, He, Min Max, Hebbring, Scott, Kaufman, Howard L, Lababidi, Samir, Lancashire, Lee J, Li, Yan, Lu, Xin X, Luo, Heng, Ma, Xiwen, Ning, Baitang, Noguera, Rosa, Peifer, Martin, Phan, John H, Roels, Frederik, Rosswog, Carolina, Shao, Susan, Shen, Jie, Theissen, Jessica, Tonini, Gian Paolo, Vandesompele, Jo, Wu, Po-Yen, Xiao, Wenzhong, Xu, Joshua, Xu, Weihong, Xuan, Jiekun, Yang, Yong, Ye, Zhan, Dong, Zirui, Zhang, Ke K, Yin, Ye, Zhao, Chen, Zheng, Yuanting, Wolfinger, Russell D, Shi, Tieliu, Malkas, Linda H, Berthold, Frank, Wang, Jun, Tong, Weida, Shi, Leming, Peng, Zhiyu *and others*. (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome biology* **16**, 133.

Zhao, Shanrong, Fung-Leung, Wai Ping, Bittner, Anton, Ngo, Karen and Liu, Xuejun. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**(1).

Zhou, Yi Hui, Barry, William T. and Wright, Fred A. (2013). Empirical pathway analysis, without permutation. *Biostatistics* **14**(3), 573–585.

Zou, Hui, Hastie, Trevor and Tibshirani, Robert. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics* **35**(5), 2173–2192.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

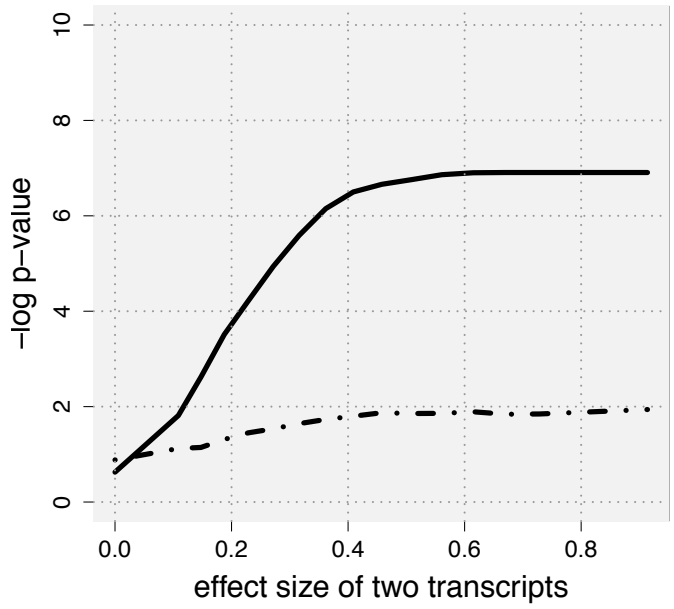

Fig. 1. Comparison of power between SAFE tests in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations. There is a practical ceiling on one curves because an empirical p-value is calculated and iterations were limited to 1000.

Table 1. *P-values from different gene-set tests on maternal peripheral blood.*

Pathway	p53	MAPK	TGF_beta	cell_prog	plac	inflamm	immune
Camera (ranks)	0.43	0.22	0.98	0.46	0.25	0.48	0.35
Camera	0.64	0.26	0.91	0.50	0.22	0.57	0.32
Roast	0.95	0.17	0.66	0.90	0.43	0.35	0.92
Safe	0.25	0.07	0.33	0.62	0.37	0.04	0.23
Gage (ranks)	0.14	0.13	0.34	0.36	0.28	0.06	0.12
Gage	0.08	0.13	0.22	0.32	0.42	0.07	0.12
JAGST comp. test	0.65	1.00	0.68	1.00	0.88	1.00	1.00
JAGST self-cont. test	0.21	1.00	0.26	0.97	0.58	1.00	1.00

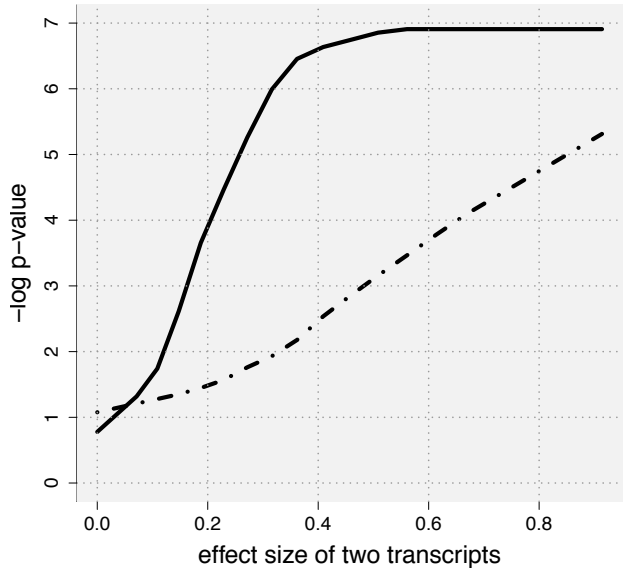

Fig. 2. Comparison of power between ROAST tests in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations. There is a practical ceiling on one curves because an empirical p-value is calculated and iterations were limited to 1000.

Table 2. *P-values from different gene-set tests on neonatal cord blood.*

Pathway	p53	MAPK	TGF_beta	cell_prog	plac	inflamm	immune
Camera (ranks)	0.76	0.69	0.61	0.83	0.38	0.54	0.67
Camera	0.93	0.71	0.66	0.88	0.45	0.59	0.70
Roast	0.84	0.93	0.63	0.77	0.66	0.81	0.91
Safe	0.47	0.55	0.65	0.85	0.31	0.12	0.75
Gage (ranks)	0.46	0.58	0.50	0.92	0.49	0.17	0.79
Gage	0.60	0.57	0.40	0.85	0.39	0.11	0.77
JAGST comp. test	0.96	1.00	0.58	1.00	0.27	1.00	1.00
JAGST self-cont. test	0.85	1.00	0.24	0.99	0.07	1.00	1.00

Table 3. *P-values from different gene-set tests on placenta data.*

Pathway	p53	MAPK	TGF_beta	cell_prog	plac	inflamm	immune
Camera (ranks)	0.10	0.86	0.45	0.27	0.09	0.34	0.98
Camera	0.10	0.95	0.49	0.21	0.07	0.20	0.91
Roast	0.23	0.33	0.78	0.41	0.18	0.04	0.43
Safe	0.18	0.03	0.15	0.35	0.75	0.11	0.31
Gage (ranks)	0.23	0.05	0.21	0.52	0.81	0.12	0.45
Gage	0.26	0.04	0.23	0.49	0.77	0.07	0.41
JAGST comp. test	0.89	1.00	0.89	1.00	0.76	1.00	1.00
JAGST self-cont. test	0.66	1.00	0.65	1.00	0.43	1.00	1.00

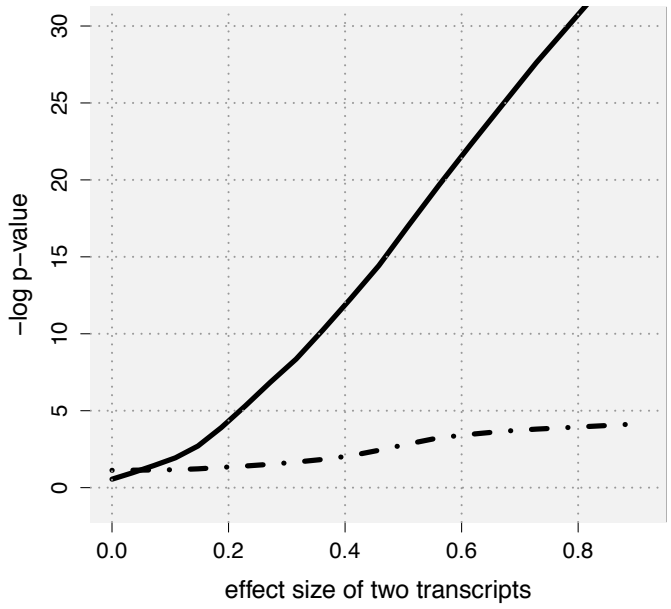

Fig. 3. Comparison of power between CAMERA tests in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations.

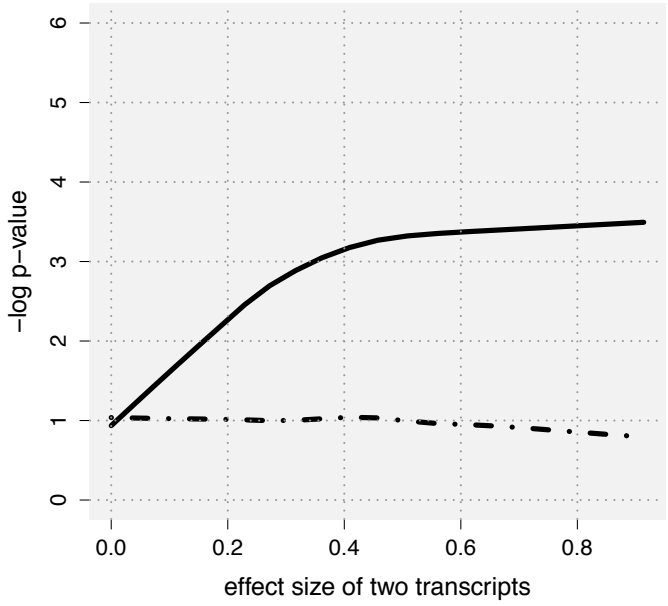

Fig. 4. Comparison of power between CAMERA tests using ranks in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations.

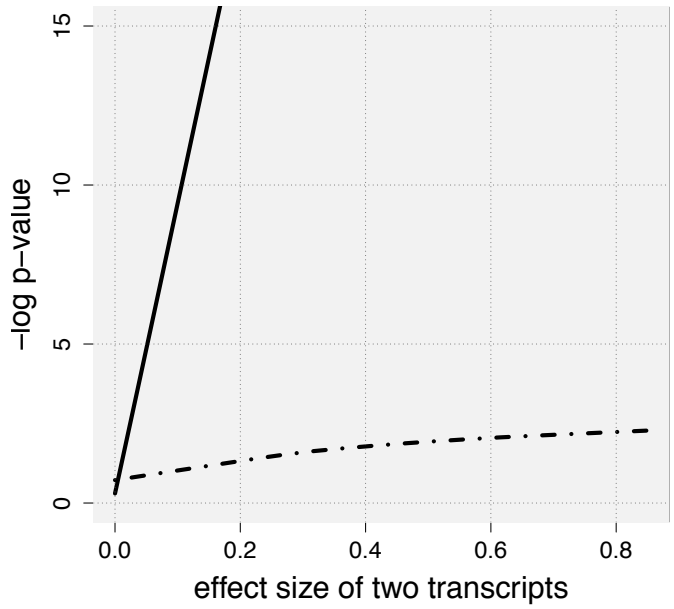

Fig. 5. Comparison of power between GAGE tests in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations.

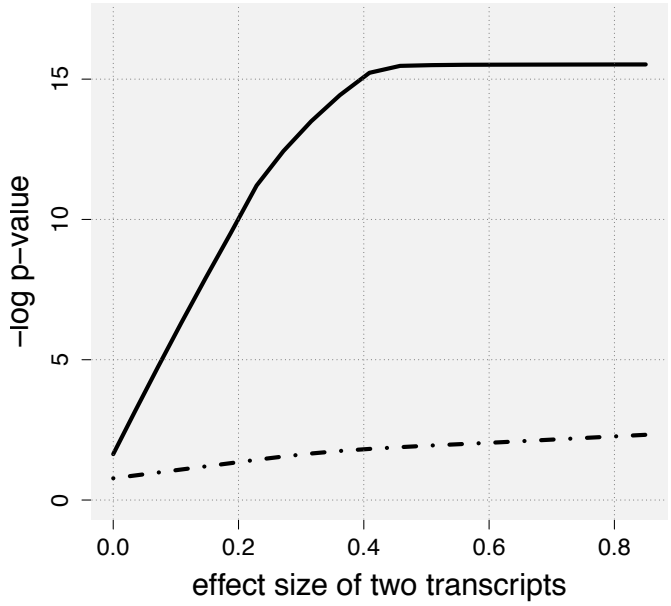

Fig. 6. Comparison of power between GAGE tests using ranks in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations.

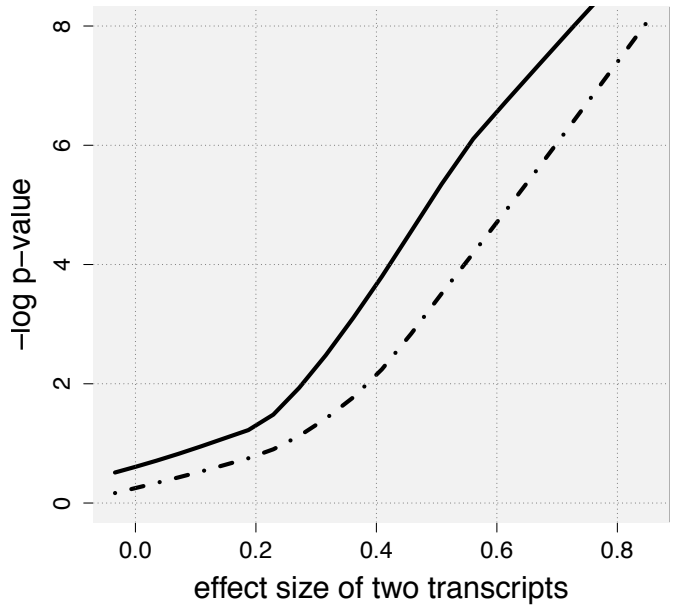

Fig. 7. Comparison of power between the JAGST competitive test in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between two transcripts within the gene set and the outcome, holding the sample size constant at 280 observations.

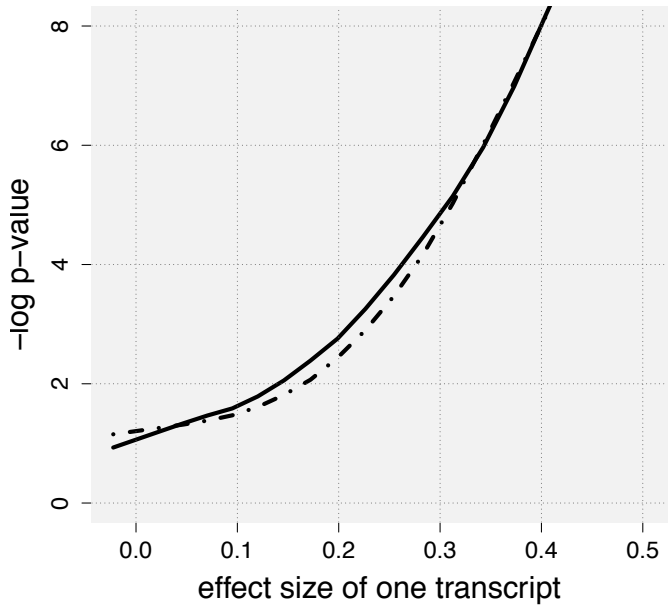

Fig. 8. Comparison of power between the JAGST self-contained test in correlated region (solid line) and uncorrelated (long dashes) region as a function of the strength of association between one transcript within the gene set and the outcome, holding the sample size constant at 280 observations.

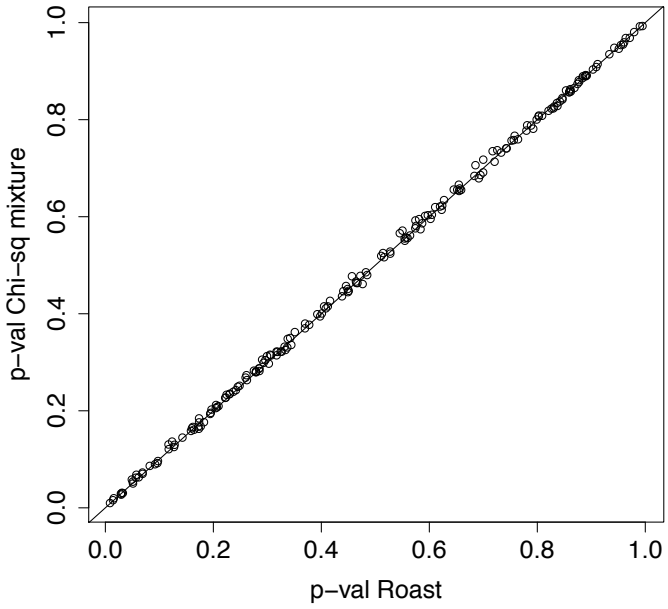

Fig. 9. Comparison of p-values between ROAST's directional test and tail probabilities calculated analytically. These p-values were generated under the null and have a correlation of 0.99. The line corresponds to y=x, along which equivalent points will fall

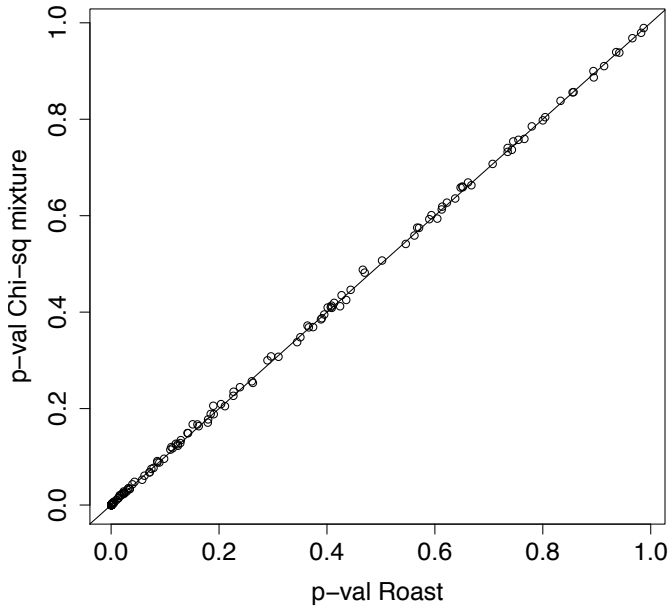

Fig. 10. Comparison of p-values between ROAST's directional test and tail probabilities calculated analytically. These p-values were generated under a weak alternative and have a correlation of 0.99. The line corresponds to y=x, along which equivalent points will fall