Competitive gene set enrichment analysis for correlated expression data

Abstract

To be filled

1 Introduction

What is enrichment analysis? Why would people care about that?
Gene set test is a method of studying the association between a set of genes, which are significantly correlated with treatment or experimental design variables, and a *prior* set of genes, which are biologically related. A typical gene expression analysis involves the detection of a set of differentially expressed genes. Differential expression (DE) analysis focuses on individual genes, and therefore it fails to provide insight into the association of treatment variable with the gene set under study. Gene set test helps researchers better understand the underlying biological processes.

What are the differences between self-contained and competitive test? And how does they work?
Depending on the definition of the null hypothesis, there are two types of gene set test: the *self-contained* and *competitive* tests (Goeman and Bühlmann, 2007). A self-contained test evaluates a set of genes by a fixed standard without reference to other genes in the genome (see Goeman et al. (2004, 2005); Tsai and Chen (2009); Wu et al. (2010); Huang and Lin (2013) for example). A competitive test compares DE genes in the test set to those not in the test set (Tian et al., 2005; Wu and Smyth, 2012; Yaari et al., 2013). Many methods, regardless of the type of test, perform a three-stage analysis (Khatri et al., 2012): on the first stage, a *gene-level statistic* that measures the association between the expression profiles and the experimental design variables is calculated for each gene; such gene-level statistics include, among others, the *signal-to-noise ratio* (Subramanian et al., 2005), the ordinary *t-statistic* (Tian et al., 2005) or a moderated t-statistic (Smyth, 2004), the *log fold change* (Kim and Volsky, 2005) and the *Z-score* (Efron, 2007). On the second stage, a *set-level statistic* is summarized by using gene-level statistics and prior information about the test set (i.e., whether the gene belongs to the test set) as input. Examples of the set-level statistics are the *enrichment score* (Subramanian et al., 2005), the *maxmean statistic* (Efron and Tibshirani, 2007), and statistic derived from joint distribution of gene-level statistics (Yaari et al., 2013), to name a few. On the last stage, a p-value is assigned to the test set by comparing the set-level statistic to its reference distribution. The competitive gene set test is much more popular among genomic literatures (Goeman and Bühlmann, 2007; Gatti et al., 2010).

Independent gene set test
Many competitive gene set test approaches rely on independence of gene-level statistics. Those tests are parametric or rank-based procedures that assume the gene-level statistics to be independent and identically distributed, or gene permutation procedures that generate the same approximate null for the set-level statistics. For example, PAGE (Kim and Volsky, 2005) conducts one-sample z-test by comparing the mean of gene-level statistics (i.e., log fold changes) in the test set to a normal distribution under the null, assuming the gene-level statistics to be independent. The 2×2 contingency-table-based tests examine the significance of the test set by dichotomizing the outcomes of DE analysis and cross-classifying the genes according to whether they are indicated as DE and whether they are in the test set (see Huang et al. (2009) for a review and references therein). sigPathway (Tian et al., 2005)

and "geneSetTest" in the limma package (Smyth, 2004) evaluate the set-level p-values by permuting gene labels. However, tests assuming independence of genes may result in inflated false discovery rate (Efron and Tibshirani, 2007; Goeman and Bühlmann, 2007; Gatti et al., 2010; Wu and Smyth, 2012; Yaari et al., 2013), as genes in a gene set are often correlated and function together.

Tests that account for between-gene correlation

A handful of methods have been proposed to account for between-gene correlation in competitive gene set test. One attempt is to evaluate the set-level statistic by permuting the biological samples (see, for example, Subramanian et al. (2005); Efron and Tibshirani (2007)). Permuting samples does not require an explicit understanding of the underlying correlation structure among genes and thus protects the test against such correlation. Since permuting sample labels is computationally inefficient, Zhou et al. (2013) proposed an analytic approximation to permutations for set-level score statistics, which preserves the essence of permutation gene set analysis with greatly reduced computational burden. However, an unavoidable problem arising from sample permutation approach is that it implicitly alters the null hypothesis being tested and it is difficult to characterize the null and the alternative hypotheses (Goeman and Bühlmann, 2007; Khatri et al., 2012; Wu and Smyth, 2012). We will further discuss this point in later sections of this paper. Another attempt is to conduct set-level test that works with the between-gene correlation structures. Wu and Smyth (2012) proposed Correlation Adjusted MEan RAnk (CAMERA) gene set test that first estimates a *variance inflation factor* (VIF) associated with correlation between gene expression profiles, and then incorporates it into two versions (i.e., the parametric and the rank-based) of CAMERA tests. Yaari et al. (2013) also used the idea of incorporating VIF to adjust for correlation in their distribution-based gene set analysis. Valid estimation of VIF relies on the assumption that correlation between any two local statistics are almost the same as correlation between their corresponding expression profiles. This assumption has been demonstrated (??? a better word???) by Barry et al. (2008) for several gene-level statistics (e.g., t-statistic, Wald-type statistic for regressing expression on censored time-to-event data through a Cox proportional hazards model). However, as shown by (the paper to be finished), this assumption holds only for the case where all of the gene-level tests are under the null (i.e., no gene is DE), and the correlation among gene-level statistics (e.g., t-statistics) can be badly estimated by sample correlation when a fraction of genes are DE.

What do we propose?

In this paper, we propose a new competitive gene set test procedure that incorporates the correlation among gene-level statistics into the set-level test statistic. This procedure aims to correct for the discrepancy between correlation among expression profiles and that among gene-level statistics in the formulation of set-level statistic. The discrepancy is caused by the presence of DE genes for several typically used gene-level statistics (REF the paper to be finished). As a remedy, our strategy is to model the covariance matrix of gene-level statistics by two variance components, one attributable to the correlation among expression profiles and the other attributable to the DE effect associated with the treatment. OurMethod follows the three-stage paradigm and works for a two group comparison experiment under all correlation structures. Our simulations show that OurMethod controls type I error correctly and maintains good power for different correlation structures we examined.

What is the plan of this paper?

The rest of the paper is organized as follows: in Section 2 we describe OurMethod.....

2 Methods

Overview of this section

In the first part of this section, we will formulate our model: first, we introduce a DE effect for each gene, based on which we derive the correlation between our gene-level statistics; then we define the null and alternative hypotheses for competitive gene set test under this framework; next we propose our set-level test statistic and conduct hypothesis testing. In the second part, we will briefly summarize four different approaches that we will compare against in the result section.

2.1 OurMethod

The DE effects

In a treatment-control gene expression experiment, we denote by Y_{ijk} a random variable for the expression level of gene i from sample j in treatment group k, with i taking the values $1, \ldots, m$ (the number of genes), j taking the values $1, \ldots, n_k$ (the total number of biological samples), and k being either 1 for control or 2 for treatment. Correspondingly, Y^*_{ijk} represents the standardized expression levels (described in REF???) for gene i of sample j, with $Y^*_{ijk} \sim N(0,1)$ (??? Normal assumption necessary here???) if sample j comes from the control group, and $Y^*_{ijk} \sim N(\Delta_i, 1)$ if it comes from the treatment group. Here, Δ_i is a *DE effect*: compared to the control group, gene i is not DE if $\Delta_i = 0$, up-regulated if $\Delta_i > 0$ and down-regulated if $\Delta_i < 0$. In a gene expression experiment, the DE effect Δ_i consists of two parts: I) the treatment which determines whether a gene is DE or not; and II) the DE effect size or strength when the gene is DE. For I), we let $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ be a vector of DE indicators, where $Z_i = 1$ if gene i is DE and $Z_i = 0$ otherwise, and (DO WE NEED TO ASSUME Z_is TO BE INDEPENDENT OF EACH OTHER?)

$$Z_i \sim \text{Binom}(1, p_i) \tag{1}$$

For II), we denote δ_i as the *DE effect size* for all genes i and δ_i follows some distribution f_δ with mean and variance

$$E(\delta_i) = \mu_\delta, \quad \text{Var}(\delta_i) = \sigma^2_\delta \tag{2}$$

We further assume that the DE indicator Z_i is independent of the DE effect size δ_i for gene $i = 1, \ldots, m$. Therefore, the DE effect can be expressed as

$$\Delta_i = Z_i \delta_i, \tag{3}$$

It can be shown that (details in Appendix 7),

$$E(\Delta_i) = p_i \mu_\delta, \quad \text{Var}(\Delta_i) = p_i \sigma^2_\delta + p_i(1-p_i)\mu^2_\delta, \quad i = 1, \ldots, m. \tag{4}$$

We assume that conditioning on the DE effects, expression levels for different samples are independent, but expression levels for different genes of the same sample may be correlated. Denote by $\boldsymbol{C}_{m\times m}$ the gene correlation matrix, with entry ρ_{i_1,i_2} being the correlation between gene i_1 and gene i_2. Note that the between-gene correlation ρ_{i_1,i_2} is a constant, regardless of whether the sample is from the treatment or from the control group. In this paper, we estimate the between-gene correlation matrix $\boldsymbol{C}$ by the residual sample correlation after the treatment differences have been nullified (as done by Efron (2007) and Wu and Smyth (2012)).

Gene-level statistics and their correlation

We denote by I_t and I_b the test set and the *background set* (i.e., the genes not in the test set). Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ be a indicator vector, with $x_i = 1$ if gene i belongs to the test set and $x_i = 0$ otherwise. Therefore $I_t = \{i : x_i = 1\}$ and $I_b = \{i : x_i = 0\}$. We assume that the DE probability is p_t for genes in the test set and p_b for genes in the background set. For gene i, the gene-level statistic is the difference in mean expression levels between the treatment and the control groups,

$$U_i = \bar{Y}_{i.2} - \bar{Y}_{i.1} \tag{5}$$

where $\bar{Y}_{i.k} = \sum_{j=1}^{n_k} Y_{ijk}/n_k$. It follows from equation (4) that $\boldsymbol{U} = (U_1, \ldots, U_m)$ has mean

$$E(U_i) = \begin{cases} p_t \mu_\delta, & \text{if } i \in I_t \\ p_b \mu_\delta, & \text{if } i \in I_b \end{cases} \tag{6}$$

and covariance matrix (see Appendix 7 for detail)

$$\text{Var}(\boldsymbol{U}) = \boldsymbol{D} + \sigma_2^2 \boldsymbol{C} \tag{7}$$

where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_m)$ with $d_i = p_t \sigma^2_\delta + p_t(1-p_t)\mu^2_\delta$ if $i \in I_t$ and $d_i = p_b \sigma^2_\delta + p_b(1-p_b)\mu^2_\delta$ if $i \in I_b$, $\sigma_2^2 = \frac{1}{n_1} + \frac{1}{n_2}$ and $\boldsymbol{C}$ is the between-gene correlation matrix.

The null hypothesis for competitive gene set test

For a competitive gene set test, it is often unclear what the hypothesed null is, and thus what is being tested (Barry et al. (2008) Wu and Smyth, 2012). Note that the DE probability affects both the mean vector in equation (6) and the covariance in equation (7). Under this framework, the test set is not enriched only if the probability of DE in the test set is the same as that in the background set. Therefore, the hypothesis for enrichment testing can be statistically formulated as

$$H_0\colon\ p_t = p_b \overset{\text{def}}{=} p_0 \text{ Versus } H_1\colon\ p_t \neq p_b \tag{8}$$

We can combine equations (6) and (7) into the following linear model

$$\boldsymbol{U} = \beta_0 \mathbf{1}_m + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon}, \quad \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{D} + \sigma_2^2 \boldsymbol{C} \tag{9}$$

with $\beta_0 = p_b \mu_\delta$, $\beta_1 = (p_t - p_b)\mu_\delta$ and $\mathbf{1}_m$ being a vector of ones. Now the hypothesis testing problem in (8) becomes

$$H_0\colon\ \beta_1 = 0 \text{ Versus } H_1\colon\ \beta_1 \neq 0. \tag{10}$$

Under the null of (10), we have $E(\boldsymbol{U}) = \beta_0 \mathbf{1}_m$ and $\text{Var}(\boldsymbol{U}) = \sigma_1^2 \boldsymbol{I}_m + \sigma_2^2 \boldsymbol{C}$ where $\boldsymbol{I}_m$ is an identity matrix and $\sigma_1^2 = p_0 \sigma_\delta^2 + p_0(1 - p_0)\mu_\delta^2$.

Set-level statistic

In practice, we need to estimate $\beta_0, \beta_1, \sigma_1^2$ and $\boldsymbol{C}$ in model (9) for gene set test. Our strategy is to use *quasi-likelihood*, which requires only the mean and the variance of $\boldsymbol{U}$. The between-gene correlation matrix $\boldsymbol{C}$ is estimated by the residual sample correlation after the treatment differences have been nullified, and is treated as known in estimating β_0 and σ_1^2. Denoting by $\hat{\boldsymbol{C}}$ the estimate of $\boldsymbol{C}$ and,

$$\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{I}_m + \sigma_2^2 \hat{\boldsymbol{C}} \tag{11}$$

The score equations for β_0 and σ_1^2 are

$$\begin{aligned} (\boldsymbol{U} - \beta_0 \mathbf{1}_m)^T \boldsymbol{\Sigma}^{-1} \mathbf{1}_m &= 0 \\ (\boldsymbol{U} - \beta_0 \mathbf{1}_m)^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{C}} (\boldsymbol{U} - \beta_0 \mathbf{1}_m) &= \text{trace}(\boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{C}}) \end{aligned} \tag{12}$$

.... something to catch up.....
The enrichment test statistic for the test set is

$$T = \frac{\left[\boldsymbol{x}^T (\boldsymbol{U} - \hat{\beta}_0 \mathbf{1}_m)\right]^2}{\left[\boldsymbol{x}^T (\boldsymbol{I} - \boldsymbol{H})\right] \boldsymbol{\Sigma} \left[\boldsymbol{x}^T (\boldsymbol{I} - \boldsymbol{H})\right]^T} \tag{13}$$

Under the null, $T \sim \chi^2(1)$.

2.2 Other competitive gene set tests

We will compare OurMethod to five existing gene set tests: Gene set enrichment analysis (GSEA, Subramanian et al. (2005)), two versions of the CAMERA procedure (Wu and Smyth, 2012), two versions of the geneSetTest procedure, and QuSAGE (Yaari et al., 2013). All tests but geneSetTest account for correlation among genes. We will denote the two versions of CAMERA by CAMERA-modt and CAMERA-rank. The first version of geneSetTest, denoted by *geneSetTest-modt*, is similar to sigPathway (Tian et al., 2005) except it uses moderated t-statistics instead of the ordinary t-statistics as gene-level statistics. The second version of geneSetTest is also known as the mean rank gene set enrichment (Michaud et al., 2008) and will be referred to herein as MRGSE. GSEA is modified from the original R-GSEA script (`http://software.broadinstitute.org/gsea/index.jsp`) to accommodate single gene set test. CAMERA and geneSetTest are implemented in the limma package (Smyth, 2005) in the Bioconductor project (Gentleman et al., 2004), and QuSAGE in the Bioconductor package of

the same name. Because GSEA and OurMethod do not support linear models, the implementations are restricted to two-group comparisons.

All but QuSAGE follow the three-stage paradigm, but the five tests are different on one or more stages. For GSEA, the gene-level statistics are the rankings of genes according to a ranking metric (we use signal-to-noise ratio, the default metric in R-GSEA), then based on the rankings an enrichment score for the test set is calculated, and the significance of the enrichment score is determined by randomly permuting the sample labels. Both CAMERA-modt and geneSetTest-modt use the moderated t-statistics (Smyth, 2004) as gene-level statistics, and determine whether the means of the gene-level statistics are significantly different for genes in the test set versus genes in the background set. The difference is how they evaluate the set-level statistics: CAMERA-modt uses a t-statistic that allows the gene-level statistics in the test set to be correlated by first estimating a variance inflation factor, and then incorporating it into the t-statistic to adjust for between-gene correlation (see materials and methods section of Wu and Smyth (2012)); geneSetTest-modt evaluates the significance of the test set by comparing the observed set-level statistics to its null distribution generated by permuting gene labels. CAMERA-rank and geneSetTest-rank conduct a Wilcoxon-Mann-Whitney rank sum test, and they amount to, respectively, CAMERA-modt and geneSetTest-modt in that they compare the rankings instead of the gene-level statistics themselves for genes in the test set to those for genes in the background set. QuSAGE generates from t-test a probability density function (PDF) for each gene, combines the individual PDFs using convolution, and quantifies gene-set activity with a complete PDF. The complete PDF can be used to compare a baseline value for self-contained gene set test, or to compare differences in expression between test set and background set in competitive gene set test.

3 Examples and Numerical Results

3.1 Simulations

In this section, we present results from type I error and power simulations under a range of between-gene correlation structures.

The simulations run as follows: first, we simulate an entire gene set containing $m = 500$ genes, from which we randomly sample $m_1 = 100$ genes to represent those in the test set, and the remaining $m_2 = 400$ genes those in the background set; second, for gene $i = 1, \ldots, m$, we simulate the DE effect Δ_i by first generating the DE size δ_i from $N(0.2, 1)$ and the DE indicator Z_i from $\text{Binom}(1, p_i)$, where $p_i = p_t$ if gene i belongs to the test set and $p_i = p_b$ otherwise, and then setting Δ_i to be the product of Z_i and δ_i; third, we set the "true" mean expression values $\mu_1 = \mathbf{0}_m$ and $\boldsymbol{\mu}_2 = \boldsymbol{\Delta}$, respectively, for the control and treatment groups; fourth, we simulate n_1 samples from $\text{MVN}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for the control group and n_2 samples from $\text{MVN}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where the covariance $\boldsymbol{\Sigma} = (\sigma_{i_1,i_2})_{m\times m}$ may take one of the following forms:

(a0): the genes are independent of each other (i.e., $\boldsymbol{\Sigma} = \boldsymbol{I}_m$).

(a): only the genes in the test set are correlated, with exchangeable correlation structure, that is, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \sigma_{i_1,i_2} = \rho$ for $\forall i_1, i_2 \in I_t$ and $\text{Cor}(Y_{i_3}, Y_{i_4}) = \sigma_{i_3,i_4} = 0$ if at least one of i_3, i_4 does not belong to I_t.

(c): all genes are correlated, with exchangeable correlation structure, that is, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \sigma_{i_1,i_2} = \rho$ for $\forall i_1, i_2 \in I$.

(e): genes are correlated within the test set and within the background set; but any two genes, one from the test set and the other from the background set, are independent. That is, the correlation structure is block diagonal, with $\text{Cor}(Y_{i_1}, Y_{i_2}) = \sigma_{i_1,i_2} = \rho_1$ for $i_1, i_2 \in I_t$, $\text{Cor}(Y_{i_3}, Y_{i_4}) = \sigma_{i_3,i_4} = \rho_2$ for $i_3, i_4 \in I_b$, and $\text{Cor}(Y_{i_5}, Y_{i_6}) = \sigma_{i_5,i_6} = 0$ for $\forall\ i_5 \in I_t, \forall\ i_6 \in I_b$.

(f): all genes are correlated, but the correlation between two genes depend on whether they belong to the test set or not. Specifically, $\text{Cor}(Y_{i_1}, Y_{i_2}) = \sigma_{i_1,i_2} = \rho_1$ for $i_1, i_2 \in I_t$, $\text{Cor}(Y_{i_3}, Y_{i_4}) = \sigma_{i_3,i_4} = \rho_2$, for $i_3, i_4 \in I_b$, and $\text{Cor}(Y_{i_5}, Y_{i_6}) = \sigma_{i_5,i_6} = \rho_3$ for $\forall\ i_5 \in I_t, \forall\ i_6 \in I_b$.

(g): genes are correlated in the same way as those from a real data.

Type I error simulations

In the above simulation setup, the test set is not enriched if DE probabilities are the same for the genes in the test set and for those in the background set (i.e., $p_t = p_b = p_0$). However, it is shown in (the paper to be finished) that the test statistics correlation between two genes is not equal to their sample correlation when at least one gene is truly DE (under two sample t-test???). Therefore, if there are true DE genes in the entire gene set, approaches assuming the same correlation between gene-level statistics and between expression values may not perform well. To illustrate this point, we performed two groups of simulations for each of the correlation structures above: in group A_1, we simulated expression data with no DE genes, that is, $p_t = p_b = 0$; and in group A_2, we simulated data sets with the same DE probabilities for all genes-specifically, DE probabilities are the same for genes in the test set and for those in the background set with $p_t = p_b = 0.2$.

For group A_1, Figure 1 shows the histograms of type I error rates for the six approaches (OurMethod, geneSetTest-modt, geneSetTest-rank, CAMERA-modt, CAMERA-rank and GSEA) under the six correlation structures. OurMethod and GSEA hold the size of type I error rates correctly for all 6 correlation structures, with simulated p-values uniformly distributed on $[0, 1]$. The two version of CAMERA control type I errors correctly for correlation structures (a0) and (a). However, both are too conservative for the case of (c) and (g), and anti-conservative for correlation structures (e) and (f). geneSetTest procedures may be too liberal depending on the underlying correlation structures.

For group A_2 simulation where DE probabilities are 0.2 across all genes, we summarize the results of the type I error rate simulation in Table **??**. OurMethod continues to hold the size of type I error rates under all six correlation structures. However, GSEA is highly skewed towards small p-values and the two versions of CAMERA procedures are too conservative under all correlation structures, and the only exception is that CAMERA controls type I error rates correctly for (a0) where genes are simulated to be independent. The two versions of geneSetTest performs reasonably well.

Explain why this happens

OurMethod shows consistent accuracy for type I error control across all simulations, but the accuracy of the other three methods may be affected by two factors: the between-gene correlation structures, and DE probability of each gene. OurMethod controls the size of type I error rates well because it takes into account the between-gene correlation and works directly with the sample correlation between genes, and therefore is robust against the two factors. (rewrite from here, because this is only my understanding.) The GSEA evaluates the enrichment score of a test set by generating its null distribution from sample permutation, and therefore the between-gene correlation is preserved when there are no DE genes, but [explain when DE exists]... For CAMERA, the set-level statistics take into account only the between-gene correlation in the test set, and therefore does not work for cases where genes in the background set are also correlated in group A_1 simulations. More importantly, according to (the paper to be finished), the variance inflation factor of the gene-level statistics (moderated t-test in Wu and Smyth (2012)) may be over-estimated when a fraction of genes are DE, and therefore the set-level test statistic is under-estimated, resulting in conservative p-values in group A_2 simulations. geneSetTest permutes the gene labels to examine the significance of the test set, and therefore it relies on independence between genes. The performances of both versions of geneSetTest are thus unpredictable in group A_1. In group A_2 where there are DE genes both in the test set and in the background set, the correlation between the gene-level statistics are smaller (in absolute value) than the correlation between the genes. Since the genes are simulated to be slightly correlated ($\rho_1 = 0.1, \rho_2 = 0.05, \rho_3 = -0.05$), the correlation between the gene-level statistics are almost negligible for geneSetTest procedure.

Figure 1: Type I error rates for gene set tests, p-value distribution for case (a0) - (g) from left to right, from top to bottom, NO gene is DE

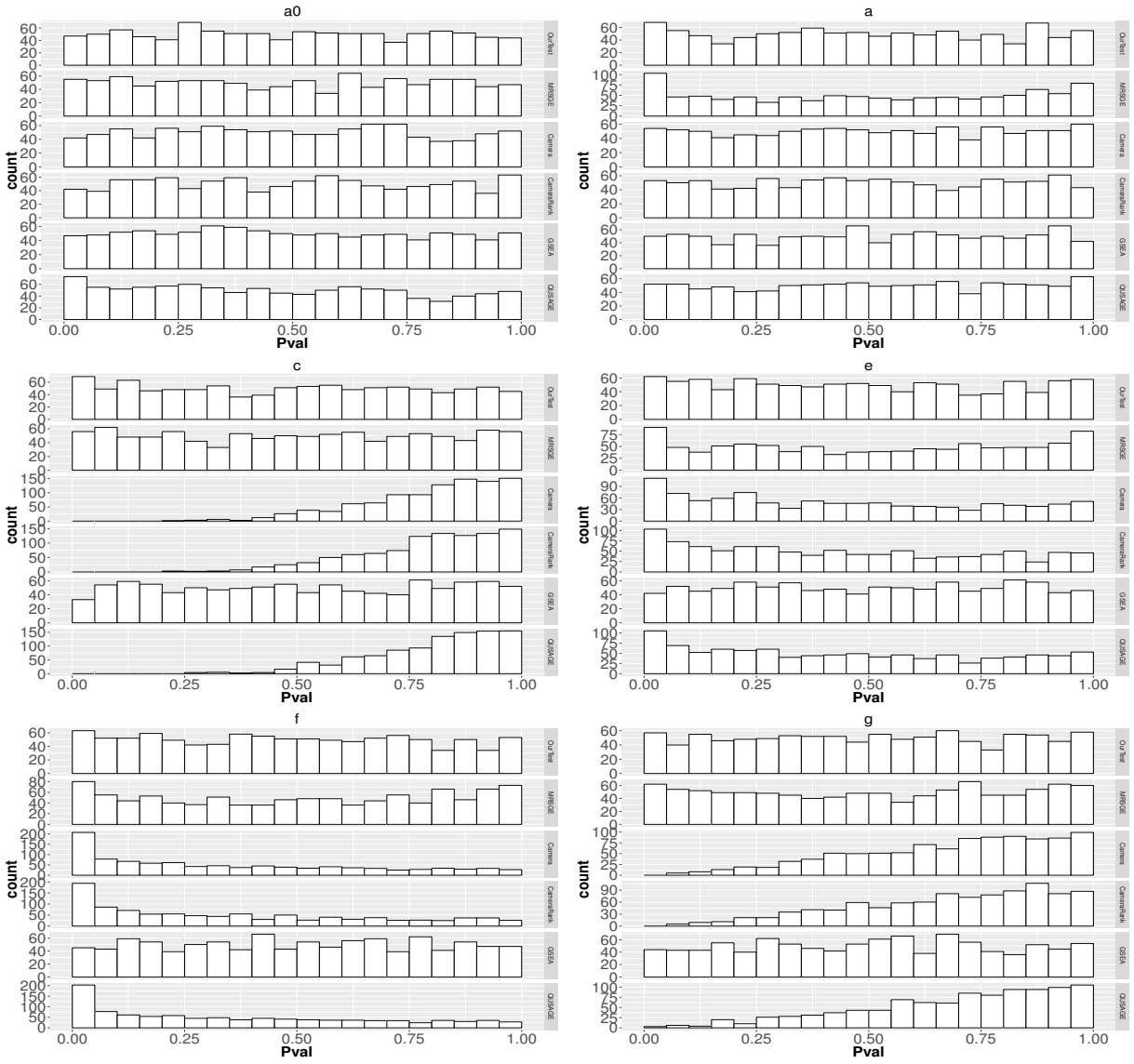
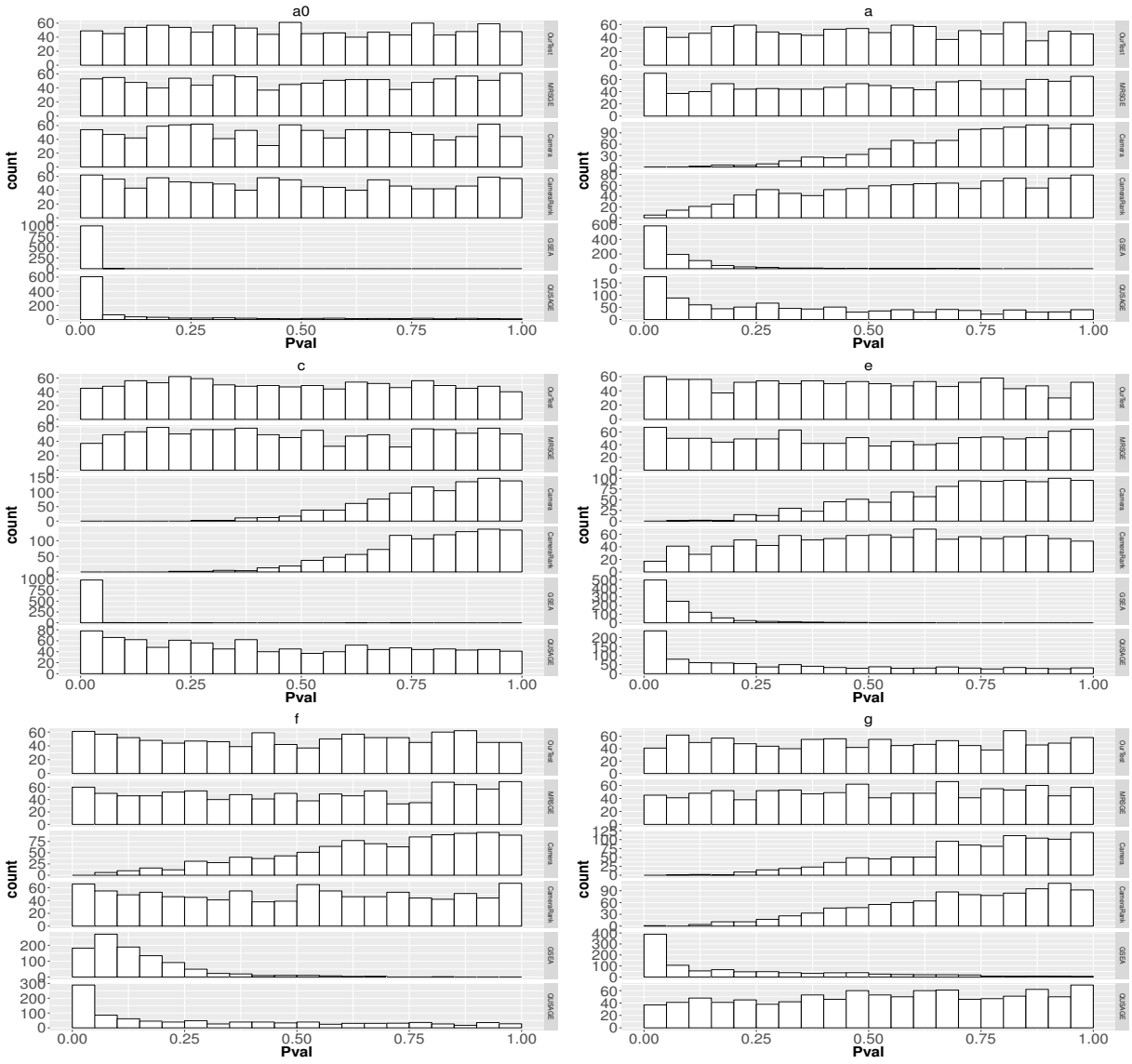


Figure 2: Type I error rates for gene set tests, p-value distribution for case (a0) - (g) from left to right, from top to bottom, DE = 20%



Power simulation

Similar to the type I error simulation, we did two groups of simulations to get a glimpse into the power issues: in group A_3, we considered four different alternative scenarios by setting DE probability to be 0% for genes in the background set and $5\%, 10\%, 15\%$ and 20% for genes in the test set; in group A_4, we set DE probability to be 10% for genes in the background set and $15\%, 20\%, 25\%$ and 30% for genes in the test set. In both groups of simulation, we fixed the DE size δ for each alternative and the six different correlation structures. The power simulation were done by generating a large number of data sets under each alternative scenario and comparing the proportion of data sets for which each test would reject at a given level α.

We compare the power of OurMethod to that of the other (How many methods) under different correlation structures. Since some of these tests are not well calibrated at the sample size considered (see Section **??**), we report calibrated power. For calibrated power, the critical value $c(\alpha)$ is chosen so that when the null hypothesis is true, exactly $100 \cdot \alpha\%$ of the resulting p-values are less than $c(\alpha)$; that is, $c(\alpha)$ is the α quantile of null distribution of p-values, where the null distribution is generated from simulation. Calibrated power allows a more fair comparison among tests, as tests that are too conservative under the null hypothesis will have greater power due to the tendency to produce small

p-values, yet this apparent power does not truly distinguish between the null and the alternative.

Table 1 and Table 2 summarize the calibrated power under correlation structure a0 (for power comparisons under the other five correlation structures, see online supplementary materials). In Table 1 we set the DE size δ to be 0.05 and simulated data in the way that genes in the background set are not DE (i.e., $p_b = 0$). GSEA has the highest calibrated power, and the rank based MRSGE has the lowest calibrated power across all four alternative scenarios. CAMERA and OurMethod have virtually no difference in the calibrated power. Furthermore, when the DE probability is 10% or higher, both CAMERA and OurMethod have comparable calibrated power to that of GSEA. In Table 2, we set the DE size δ to be 0.1 and simulated data with 10% of the genes in the background set to be DE. CAMERA and OurMethod still have indistinguishable calibrated power and both are better than MRSGE. GSEA has virtually no power at all. These results indicate that....

Table 1: power simulation results

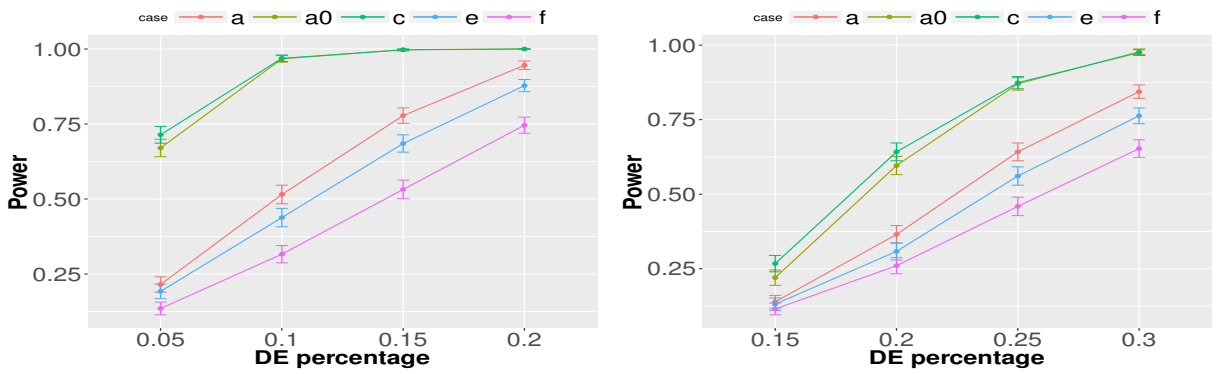
Method	DE size = 0.05			
	5% VS 0%	10% VS 0%	15% VS 0%	20% VS 0%
OurTest	0.660(0.015)	0.967(0.006)	0.998(0.001)	1.000(0.000)
lm	0.674(0.015)	0.970(0.005)	0.997(0.002)	1.000(0.000)
ModeratedT	0.856(0.011)	0.994(0.002)	0.999(0.001)	1.000(0.000)
MRSGE	0.201(0.013)	0.417(0.016)	0.682(0.015)	0.901(0.009)
Camera	0.663(0.015)	0.964(0.006)	0.997(0.002)	1.000(0.000)
CameraRank	0.133(0.011)	0.297(0.014)	0.583(0.016)	0.828(0.012)
GSEA	0.847(0.011)	0.994(0.002)	1.000(0.000)	1.000(0.000)
QUSAGE	0.747(0.014)	0.984(0.004)	0.999(0.001)	1.000(0.000)

Table 2: power simulation results

Method	DE size = 0.1			
	15% VS 10%	20% VS 10%	25% VS 10%	30% VS 10%
OurTest	0.243(0.015)	0.562(0.006)	0.892(0.001)	0.980(0.000)
lm	0.245(0.015)	0.578(0.005)	0.897(0.002)	0.984(0.000)
ModeratedT	0.325(0.011)	0.670(0.002)	0.921(0.001)	0.991(0.000)
MRSGE	0.176(0.013)	0.413(0.016)	0.716(0.015)	0.898(0.009)
Camera	0.254(0.015)	0.564(0.006)	0.891(0.002)	0.982(0.000)
CameraRank	0.130(0.011)	0.272(0.014)	0.620(0.016)	0.810(0.012)
GSEA	0.000(0.011)	0.000(0.002)	0.000(0.000)	0.000(0.000)
QUSAGE	0.265(0.014)	0.589(0.004)	0.899(0.001)	0.987(0.000)

Figure 3 shows for OurMethod, the variations in power according to different correlation structures across four alternative scenarios. The left part is the power for group A_3 with a DE size of 0.05, and the right part is the power for group A_4 with a DE size of 0.05. For each correlation structure and each alternative, we simulated 1000 data sets and calibrated the power at a significance level of 0.05. The powers for case (a0) and (c) are very similar, and are among the highest under each of the four alternatives. It's not surprising because they correspond to the simplest correlation structures: gene expression values in (a0) are simulated to be independent and in (c) are simulated to have the same correlation 0.1. As the correlation structure becomes more complicated, from (a) to (e) then to (f), the power decreases under every alternative scenario. The power under correlation structure (f) is the lowest for both A_3 and A_4 simulations. We also note that the power decreases when DE genes are present in the background set as compared to the case where there are no DE genes in the background set. (We might need the same DE size to illustrate this point.)

Figure 3: power: left, DE size = 0.05, and no DE genes in the background set; right; DE size = 0.05, DE probability is 0.1 in the background set.



3.2 Maybe real data analysis???

4 Discussion

There are many methods developed for gene set tests (see reviews by Huang et al. (2009); Khatri et al. (2012); Tarca et al. (2013)). Using the terminology of Khatri et al. (2012), these methods generally fall into three categories: *over-representation analysis*, *functional class scoring* and *pathway topology*. The over-representation analysis evaluates a fraction of genes among a set of pre-selected interesting genes (e.g., differentially expressed genes between treatment versus control samples). The test is usually conducted in the form of 2×2 table, for example, GOstat of Klebanov et al. (2007) and GO:TermFinder of Tian et al. (2005). However, the over-representation analysis methods have inherent limitations such as information loss by choosing arbitrary threshold (e.g., p-value < 0.05), or problematic assumption of independence of genes (Goeman and Bühlmann (2007); Wu and Smyth (2012)). The functional class scoring performs three-stage analysis (Khatri et al., 2012): on the first stage, a *gene-level statistic* that measures the association between the expression profiles and the experimental design variables is calculated for each gene; such gene-level statistics include, among others, signal-to-noise ration (Subramanian et al., 2005), moderated t statistics (Smyth, 2004) and Z-score (Efron, 2007). On the second stage, a *set-level statistic* is calculated by using gene-level statistics and prior information about the test set (i.e., whether the gene belongs to the set) as input. On the last stage, a p-value is assigned to the test set by comparing the set-level statistic to its reference distribution. (Rewrite this part)

The pathway topology will not be discussed in this paper (Khatri et al. (2012); Tarca et al. (2013))

5 Conclusion

6 AcknowledgeMents

7 Appendix

First $E(\Delta_i) = E(Z_i\delta_i) = E(Z_i)E(\delta_i) = p_i\mu_\delta$. Next note that

$$\mathrm{Var}(\Delta_i) = E[(Z_i\delta_i)^2] - [E(Z_i\delta_i)]^2 = \mathrm{Var}(Z_i)[E(\delta_i)]^2 + \left[(EZ_i)^2 + \mathrm{Var}(Z_i)\right]\mathrm{Var}(\delta_i) = p_i\sigma_\delta^2 + p_i(1-p_i)\mu_\delta^2$$

Let $T_i = \bar{Y}_{i,2} - \bar{Y}_{i,1}$ be the difference in mean expression levels between the treatment group and the control group. We have

$$E(T_i) = E(\bar{Y}_{i,2}) - E(\bar{Y}_{i,1}) = E(\Delta_i) = E(Z_i\delta_i) = p_i\mu_\delta$$

The covariance between two genes i_1 and i_2 is given by (I HAVE CONCERNS HERE, IS IT VALID TO ASSUME THAT DE EFFECTS ARE INDEPENDENT BETWEEN GENES? WE SEE CO-EXPRESSION!! OR WE'VE ALREADY TAKEN THAT INTO ACCOUNT BY CORRELATION BETWEEN GENES"),

$$\begin{aligned}\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E\left[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})\right] + \mathrm{Cov}\left[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})\right] \\ &= E\left(\frac{1}{n_1}\rho_{i_1,i_2} + \frac{1}{n_2}\rho_{i_1,i_2}\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}\end{aligned} \tag{14}$$

For gene i, the variance $\mathrm{Var}(T_i) = \mathrm{Var}(\bar{Y}_{i,1}) + \mathrm{Var}(\bar{Y}_{i,2})$, with

$$\mathrm{Var}(\bar{Y}_{i,1}) = \frac{1}{n_1}$$

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2})\right] \\ &= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}) \\ &= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\ &+ \frac{n_2-1}{n_2}\left[E\left(\mathrm{Cov}(Y_{ij_1 2}, Y_{ij_2 2}|\Delta_i)\right) + \mathrm{Cov}\left(E(Y_{ij_1 2}|\Delta_i), E(Y_{ij_2 2}|\Delta_i)\right)\right] \\ &= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)\end{aligned} \tag{15}$$

Therefore $\mathrm{Var}(T_i) = \frac{1}{n_1} + \frac{1}{n_2} + \mathrm{Var}(\Delta_i)$, and it follows that

$$\mathrm{Cov}(\boldsymbol{T}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{16}$$

where $\boldsymbol{D}$ is a diagonal matrix with $\mathrm{Var}(\Delta_i) = p_i\sigma_\delta^2 + p_i(1-p_i)\mu_\delta^2$ as its ith diagonal element, and $\sigma_2^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

References

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

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477).

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The annals of applied statistics*, pages 107–129.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC genomics*, 11(1):574.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

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

Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K., and Van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957.

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

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13.

Huang, Y.-T. and Lin, X. (2013). Gene set analysis using variance component tests. *BMC bioinformatics*, 14(1):210.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375.

Kim, S.-Y. and Volsky, D. J. (2005). Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, 6(1):144.

Klebanov, L., Glazko, G., Salzman, P., Yakovlev, A., and Xiao, Y. (2007). A multivariate extension of the gene set enrichment analysis. *Journal of bioinformatics and computational biology*, 5(05):1139–1153.

Michaud, J., Simpson, K. M., Escher, R., Buchet-Poyau, K., Beissbarth, T., Carmichael, C., Ritchie, M. E., Schütz, F., Cannon, P., Liu, M., et al. (2008). Integrative analysis of runx1 downstream pathways and target genes. *BMC genomics*, 9(1):363.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3, Article3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549.

Tsai, C.-A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, 25(7):897–903.

Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. (2010). Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182.

Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133.

Yaari, G., Bolen, C. R., Thakar, J., and Kleinstein, S. H. (2013). Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic acids research*, page gkt660.

Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation. *Biostatistics*, page kxt004.