Biometrics 000, 000–000
000 0000

DOI: 000

Accounting for correlations in gene set test for improved interpretation of genome-scale data

Summary: Functional enrichment analysis is a widely used tool for interpreting high-throughput biological data, such as gene expression and proteomics data. It aims at testing categories of genes for enriched association signals in a list of genes inferred from genome-wide data. Most conventional enrichment testing methods ignore or do not properly account for the widespread correlations among genes, which, as we show, can result in severely inflated type 1 error rates and power loss. We propose a new framework for enrichment testing based on a mixed effects quasi-likelihood model, where the data are not required to be Gaussian. Our method effectively adjusts for completely unknown, unstructured correlations among the genes. It uses a score test approach and allows for analytical assessment of p-values, which makes it computationally rapid for massive omics data. Compared to existing methods including GSEA and CAMERA, our method enjoys robust and substantially improved control over type 1 error and maintains good power in a variety of correlation structure and association settings. We also present a real data analysis to illustrate our approach.

Key words:

1. Introduction

What is enrichment analysis? Why would people care about that?

Gene set test is a statistical framework of studying the association between a test set—a *prior* set of biologically related genes—and a set of genes that are significantly correlated with treatment or experimental design variables. A key task of gene expression analysis involves the detection of a set of differentially expressed genes. Differential expression (DE) analysis evaluates each individual gene separately, and therefore it fails to provide insight into the relation between treatment variables and the prior gene set under study. Gene set test helps researchers better understand the underlying biological processes in terms of ensembles of genes.

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* test and the *competitive* test (Goeman and Bühlmann, 2007). A self-contained test examines a set of genes by a fixed standard without reference to other genes in the genome (see, for example, Goeman et al. (2004, 2005); Tsai and Chen (2009); Wu et al. (2010); Huang and Lin (2013)). 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* is calculated for each gene in the whole genome to measure the association between the expression profiles and the experimental design variables; such gene-level statistic includes, among others, *signal-to-noise ratio* (Subramanian et al., 2005),*ordinary t-statistic* (Tian et al., 2005) or *moderated t-statistic* (Smyth, 2004), *log fold change* (Kim and Volsky, 2005) and *Z-score* (Efron, 2007). On the second stage, a *set-level statistic* is obtained by utilizing the gene-level statistics from the first stage and their membership with respect to the test set (i.e., whether the gene belongs to the test set). Examples of the set-level statistic are *enrichment score* (Subramanian et al., 2005), *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 third 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 tests rely on independence of gene-level statistics that further depends on independence among genes. 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., the mean of log fold changes) in the test set to a normal distribution under the null. 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 within a gene set are often co-expressed 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 sample labels (see, for example, Subramanian et al. (2005); Efron and Tibshirani (2007)). Permuting sample labels 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 therefore difficult to characterize the null and the alternative hypotheses (Goeman and Bühlmann, 2007; Khatri et al., 2012; Wu and Smyth, 2012). Another attempt is to use set-level statistic that directly includes between-gene correlation estimated from the data. For example, CAMERA (Wu and Smyth, 2012) calculates a *variance inflation factor* (VIF) from sample correlation (after the treatment effect removed), and then incorporates it into the two versions (i.e., the parametric and the rank-based) of set-level statistics. Yaari et al. (2013) also used the idea of estimating the VIF to adjust for correlation in their distribution-based gene set test. The VIF is a crucial factor and valid estimation of it 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 simulation (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 the hypotheses 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?

We propose a new framework, need_a_name, for gene set test based on a mixed effects quasi-likelihood model. Our strategy is to avoid the discrepancy between correlations among expression profiles and those among gene-level statistics caused by the presence of DE genes. Instead, we use differences in mean as gene-level statistics for a two group comparison experiment. We model the covariance of gene-level statistics by two variance components, one attributable to correlations among samples after treatment effect removed, and the other attributable to the DE effect associate with the treatment. The benefit of quasi-likelihood is that the data are not required to be Gaussian. Our method effectively adjusts for completely unknown, unstructured correlations among the genes. need_a_name uses a score test approach and allows for analytical assessment of p-values. Compared to existing methods including GSEA and CAMERA, need_a_name enjoys robust and improved control over type I error and maintains good power in a variety of correlation structure and association settings.

What is the plan of this paper?
The rest of the paper is organized as follows: in Section 2 we describe the methodology of need_a_name; in Section 3 we present results from comparison of need_a_name to other existing methods by simulation study, and illustrate the application of our method by two real data sets; in Section 4 we conclude and also specifies the future work.

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 *need_a_name*

2.1.1 *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 6),

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

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)).

2.1.2 *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 6 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 $d_i = p_t\sigma_\delta^2 + p_t(1-p_t)\mu_\delta^2$ if $i \in I_t$ and $d_i = p_b\sigma_\delta^2 + p_b(1-p_b)\mu_\delta^2$ 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.

2.1.3 *The null hypothesis for competitive gene set test.* For a competitive gene set test, it is often unclear what the hypothesized 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

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

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

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

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

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

Under the null of (10), we have $E(\boldsymbol{U}) = \beta_0\boldsymbol{1}_m$ 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 $\sigma_1^2 = p_0\sigma_\delta^2 + p_0(1-p_0)\mu_\delta^2$.

2.1.4 *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

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

.... something to catch up.....

The enrichment test statistic for the test set is

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

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

2.2 *Other competitive gene set tests*

We will compare need_a_name to six 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 is available in the Bioconductor package of the same name. Because GSEA and need_a_name do not support linear models, the implementations are restricted to two-group comparisons.

The six tests differ in one or more aspects, although all tests except QuSAGE follow the three-stage paradigm described in Section 1. 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 throughout this paper), 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 VIF, 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 accesses 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 MRGSE 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 profiles 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. Since a standardization procedure is required by need_a_name for preprocessing data, we will simulate the standardized expression profiles for illustration purpose.

Let Y_i be the expression profile of gene i and $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_{i_1,i_2}$ for any two genes i_1 and i_2. We assume that $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_1$ if genes i_1 and i_2 are both from the test set (i.e., $i_1, i_2 \in I_t$ where I_t denote the test set), $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_2$ if they are both from the background set (i.e., $i_1, i_2 \in I_b$ where I_b denote the background set), and $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_3$ if i_1 is from the test set and i_2 is from the background set (i.e., $i_1 \in I_t, i_2 \in I_b$). We examine five different correlation structures, listed as follows:

(a): $\rho_1 = \rho_2 = \rho_3 = 0$; that is, the genes are independent of each other.

(b): $\rho_1 = 0.1$, $\rho_2 = \rho_3 = 0$; that is, only the genes in the test set are correlated.

(c): $\rho_1 = \rho_2 = \rho_3 = 0.1$; that is, all genes are correlated, with an exchangeable correlation structure.

(d): $\rho_1 = 0.1$, $\rho_2 = 0.05$, $\rho_3 = 0$; that is, 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.

(e): $\rho_1 = 0.1$, $\rho_2 = 0.05$, $\rho_3 = -0.05$; that is, all genes are correlated, but the correlation between two genes depend on whether they belong to the test set or not.

The simulations run as follows: first, we generate 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 set the DE size δ_i to be 0.1 and simulate the DE indicator Z_i from $\mathrm{Binom}(1, p_i)$, where $p_i = p_t$ if gene i belongs to the test set and $p_i = p_b$ otherwise, and then the DE effect Δ_i is the product of Z_i and δ_i; third, we set the "true" mean expression values $\boldsymbol{\mu}_1 = \mathbf{0}_m$ and $\boldsymbol{\mu}_2 = \boldsymbol{\Delta}$, respectively, for the control and treatment groups; fourth, we simulate n_1 samples from $\mathrm{MVN}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for the control group and n_2 samples from $\mathrm{MVN}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for the treatment group, where the covariance $\boldsymbol{\Sigma} = [\mathrm{Cov}(Y_{i_1}, Y_{i_2})]_{m \times m}$ may be one of the correlation structures in (a)-(e).

We have mentioned in the Introduction part that the test statistics correlations among genes are not equal to their sample correlations 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 almost equality of correlations among gene-level statistics and those among expression values may not perform well. To illustrate this point, we performed two groups of simulations for each of (a)-(e) correlation structures. In both type I error and power simulations, we set the DE probability to be 0% in group A_1 and 10% in group A_2 for genes in the background set. In the type I error simulation, we have $p_t = p_b$ under the null. In the power simulation, we considered four different alternative scenarios S_1-S_4: for genes in the test set, we set DE probability to be $5\%, 10\%, 15\%$ and 20% in group A_1, and $15\%, 20\%, 25\%$ and 30% in group A_2. Table 1 summarizes the simulation setup for the two groups.

Table 1: Parameter setup for type I error and power simulations. S_1-S_4 represent the four alternatives regarding power simulation.

Group	p_b	p_t			
		S_1	S_2	S_3	S_4
A_1	0%	5%	10%	15%	20%
A_2	10%	15%	20%	25%	30%

p_b: DE probablility for genes in the background set.
p_t: DE probablility for genes in the test set.

3.1.1 *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., $p_t = 0\%$ for group A_1 and $p_t = 10\%$ for group A_2). We expect some tests to have different performances between group A_1 and A_2 simulations under certain correlation structures.

We report the type I error simulation results for group A_1 and A_2 simulations. Figure 1 shows the uniform quantile-quantile (QQ) plots of p-values for the seven approaches (need_a_name, geneSetTest-modt, MRSGE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) under each of the five correlation structures. The left column of Figure 1 shows the p-values of group A_1 simulations. Each plot, from top to bottom, corresponds accordingly to correlation structures (a)-(e). GSEA and need_a_name hold the size of type I error rate correctly for all five correlation structures, with simulated p-values uniformly distributed on $[0, 1]$. The two version of CAMERA and QuSAGE control type I errors correctly for correlation structures (a) and (b), yet they are too conservative for the case of (c) and anti-conservative for correlation structures (d) and (e). geneSetTest-modt and MRSGE procedures have well-calibrated type I error for correlation structures (a) and (c), but are biased towards both smaller and larger p-values for the case of (b), (d) and (e).

The right column of Figure 1 shows type I error rate of

group A_2 simulations. need_a_name continues to hold the size of type I error rate, whereas GSEA is highly skewed towards small p-values, under all five correlation structures. The two versions of CAMERA control type I error rate correctly for (a) where genes are simulated to be independent. CAMERA-modt is too conservative under (b)–(e) , and CAMERA-rank may be liberal (conservative in (b) and (c), and anti-conservative in (e)). QuSAGE is too conservative under all correlation structures and the only exception is that it's anti-conservative in (c). The two versions of geneSetTest have similar performance as they do in A_1 simulations except the resulting p-values are less biased.

Explain why this happens

need_a_name shows consistent accuracy for type I error control across all simulations, but the accuracy of the other six methods may be affected by two factors: the between-gene correlation structures, and DE probability of each gene. need_a_name controls the size of type I error well because it uses difference in mean as gene-level statistic, and the correlations among gene-level statistics are fully reflected in sample correlation. GSEA evaluates the enrichment score of a test set by generating its null distribution from sample permutation. When there's no DE genes such as in the case of group A_1 simulation, GSEA performs extremely well since permuting sample labels won't change the underlying correlation structure. When DE genes exist, however, sample permutation will destroy the between-gene correlation structure, which explains the complete failure of GSEA in controlling type I error for the case of group A_2 simulation. For CAMERA and QuSAGE, according to (the paper to be finished), the VIF 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. The performances of those methods—two versions of CAMERA and QuSAGE—are subject to the underlying correlation structures. Moreover, the performance of CAMERA is complicated by the fact that the set-level statistic takes into account only the between-gene correlation in the test set without addressing that in the background set.

Different from the five methods mentioned above, geneSetTest-modt and MRSGE rely on independence between genes. It's not surprising that gene permutation method, such as geneSetTest-modt and MRSGE, controls type I error correctly when genes are "equally-correlated": in (a) genes are simulated to be independent, and in (c) genes are simulated to have an exchangeable correlation structure. However, both geneSetTest-modt and MRSGE fail to hold type I error size for the remaining three correlation structures. We note that both methods perform better in group A_2 as compared to their counterpart in group A_1 simulation under each of (b), (d) and (e) correlation structures. 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 is smaller (in absolute value) than between sample correlation. 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 to work.

3.1.2 *Power simulation.* The power simulation were done by generating $10,000$ data sets under each alternative scenario S_1-S_4 (see Table 1) and comparing the proportion of data sets for which each test would reject at a given level α.

We compare the power of need_a_name to those of the other six methods under different correlation structures. Since some of these tests are not well calibrated at the sample size considered (see results in Section 3.1.1), 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 2 summarizes the calibrated power for the two groups of simulations (i.e., A_1 and A_2 in Table 1). We only report the results for correlation structure (a) where genes are simulated to be independent (for power comparisons under the other four correlation structures, see online supplementary materials...). For A_1 simulations, GSEA and geneSetTest-modt have the highest, and rank based methods MRSGE and CAMERA-rank have the lowest, calibrated power across all four alternative scenarios (the data for S_4 not shown). CAMERA-modt and need_a_name have virtually no difference in the calibrated power. Furthermore, when the DE probability is 10% or higher (i.e., the case of S_2-S_4), both CAMERA-modt and need_a_name have comparable calibrated power to that of GSEA and geneSetTest-modt. In group A_2, geneSetTest-modt continues to achieve the highest calibrated power while GSEA shows virtually no power. CAMERA-modt and need_a_name still have indistinguishable calibrated power and both are better than MRSGE and CAMERA-rank.

Figure 2 shows for need_a_name, the variations in power according to different correlation structures across four alternative scenarios S_1-S_4. For each correlation structure and each alternative, we report the power (without recalibration) at a significance level of 0.05. The top is the power for group A_1, and the bottom for group A_2. The powers for case (a) 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 (a) are simulated to be independent and in (c) are simulated to have the same correlation 0.1. As the correlation structure becomes more complex, from (b) to (d) then to (e), the power decreases under every alternative scenario. The power under correlation structure (e) is the lowest for both A_1 and A_2 simulations.

3.2 *Real Data*

We applied need_a_name to two example data sets, and compared the enriched gene sets to those obtained by GSEA and by CAMERA-modt. In both examples, need_a_name were able to identify more gene sets as enriched. Our results lend credence to previous studies in finding potential gene sets

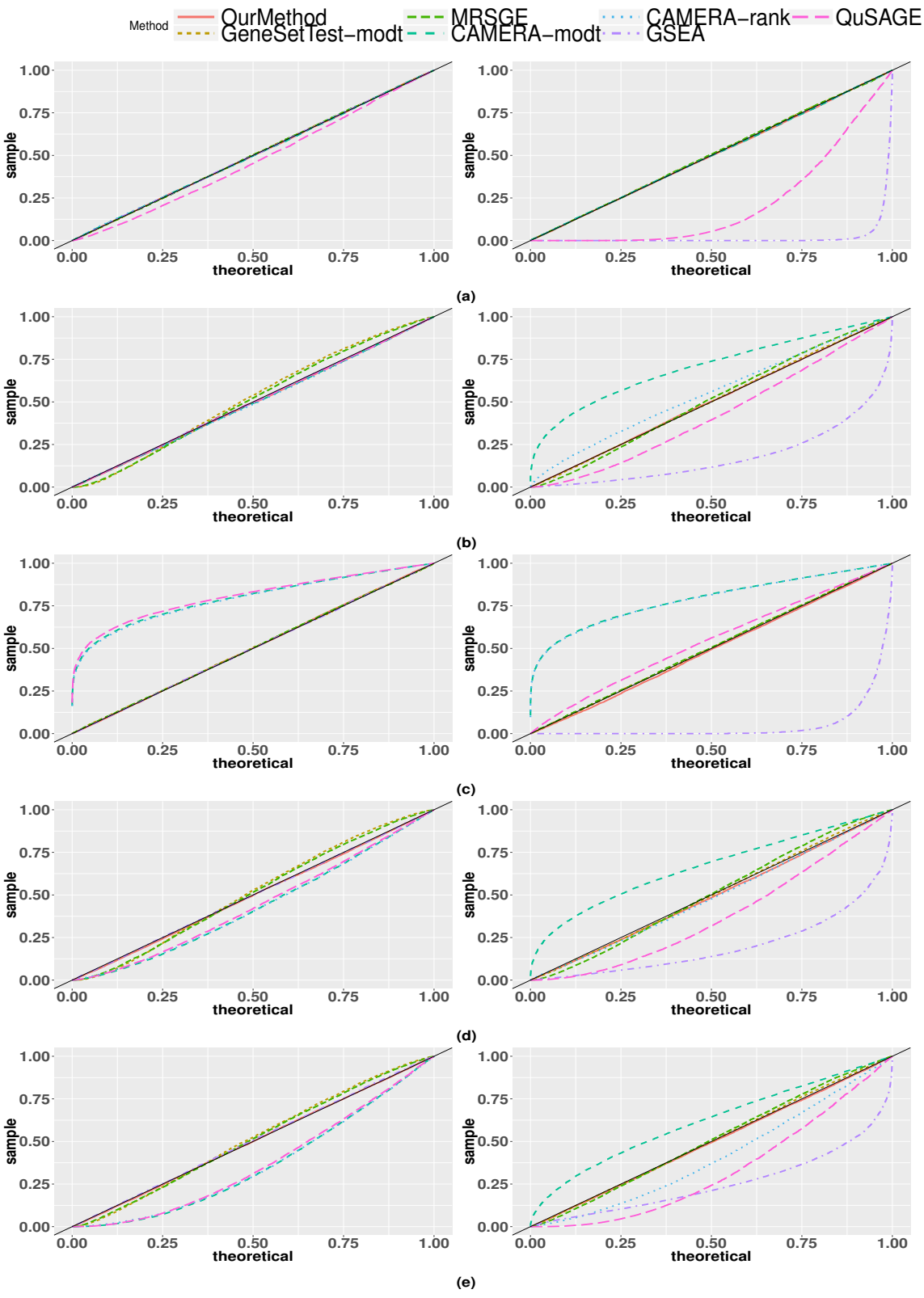

Figure 1: Uniform quantile-quantile plots for p-values by different methods. Each plot from top to bottom corresponds to correlation structures (a)-(e), respectively. The left column is for group A_1 simulation, and the right column for group A_2 simulation (see Table 1 for detail). Results are based on 10,000 simulations.

Table 2: Recalibrated power (standard error) for different methods. The powers are summarized under three alternatives S_1-S_3 in each of the group A_1 and A_2 simulations (see Table 1 for detail). Results are based on 10,000 simulations.

Method	Group A_1			Group A_2		
	S_1	S_2	S_3	S_1	S_2	S_3%
need_a_name	0.654(0.005)	0.956(0.002)	0.998(0.000)	0.229(0.004)	0.604(0.005)	0.871(0.003)
geneSetTest-modt	0.825(0.004)	0.989(0.001)	1.000(0.000)	0.322(0.005)	0.704(0.005)	0.920(0.003)
MRSGE	0.186(0.004)	0.426(0.005)	0.701(0.005)	0.183(0.004)	0.423(0.005)	0.700(0.005)
CAMERA-modt	0.647(0.005)	0.953(0.002)	0.998(0.000)	0.227(0.004)	0.596(0.005)	0.864(0.003)
CAMERA-rank	0.126(0.003)	0.324(0.005)	0.585(0.005)	0.113(0.003)	0.310(0.005)	0.570(0.005)
GSEA	0.827(0.004)	0.991(0.001)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
QuSAGE	0.723(0.004)	0.974(0.002)	0.999(0.000)	0.244(0.004)	0.630(0.005)	0.889(0.003)

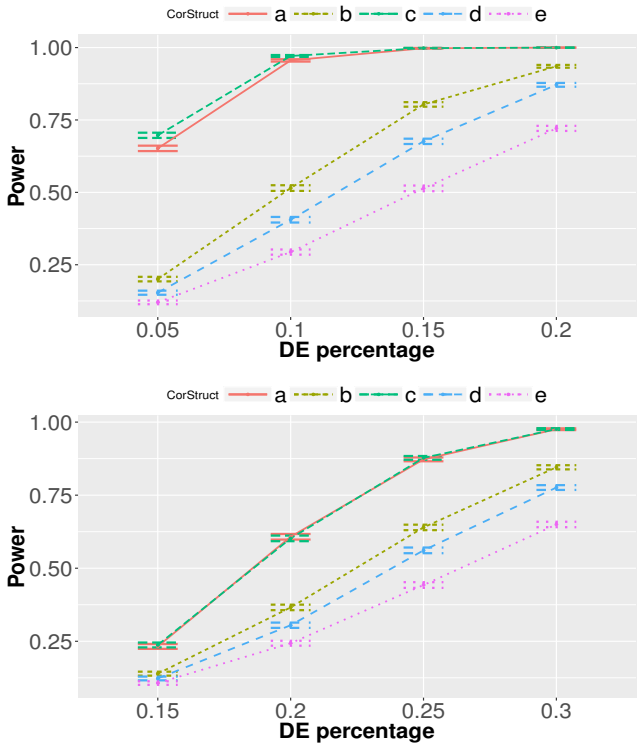

Figure 2: Power for need_a_name under correlation structures (a)-(e) of Section 3.1. The top corresponds to group A_1 simulations, and the bottom to group A_2 simulations (see Table 1). The error bars are the 95% CIs based on 10,000 simulations.

correlated with Huntington's disease and those correlated with chromosome Y and Y bands in lymphoblastoid cells.

3.2.1 *Huntington's Disease Data.* We examined the Huntington's Disease (HD) RNA-Sequencing (RNA-Seq) data (Labadorf et al., 2015) to identify which gene sets are enriched among DE genes in HD. The mRNA expression profiles in human prefrontal cortex were obtained from 20 Huntington's Disease samples and 49 neurologically normal controls. Expression values were normalized and filtered as described in the methods section of Labadorf et al. (2015). The data, containing 28,087 genes, is available as a series GSE64810 in the GEO database (`http://www.ncbi.nlm.nih.gov/geo/`). We performed enrichment analysis using the MsigDB (Subramanian et al., 2005) C2 Canonical Pathways gene sets (February 5, 2016, data last accessed). The C2 Canonical Pathway gene sets have a collection of 1330 gene sets, with an average set size of 50 (the set sizes range from 3 to 1028, and the median is 29). Since the genes in C2 are named by HGNC symbols and by ensembl IDs in the HD expression data set, we converted the ensembl IDs in the expression data into HGNC symbols using *BioMart* (`http://uswest.ensembl.org/biomart/martview/`). We retained 26,941 genes that have corresponding HGNC symbols. For need_a_name, we standardized the data in the way as described in Section....

We used three test procedures (need_a_name, GSEA and CAMERA-modt) to run enrichment analysis for the entire C2 Canonical Pathway gene sets, and compared the three tests in terms of resulting enriched gene sets. need_a_name found 176 out of 1330 gene sets to be enriched using the Benjamini-Hochberg (BH) procedure at a false discovery rate (FDR) of 0.05 (for multiple hypothesis testing, unless specified otherwise, all p-values in Section 3.2 were adjusted by BH procedure). GSEA found 9 enriched gene sets—8 of them were also among the 176 gene sets we identified (the one that was not significant according to need_a_name had a p-value of 0.008 and FDR 0.057). CAMERA-modt found no enriched gene sets. In Figure 3 we present pairwise p-value plots for need_a_name, GSEA and CAMERA-modt. When plotted against p-values of GSEA, for need_a_name, smaller p-values (e.g., less than 0.1) are more likely to cluster—as compared to larger p-values; that is, need_a_name produces more small p-values than GSEA does while need_a_name and GSEA do not differ much in producing larger p-values. The p-values of CAMERA-modt are overwhelmingly larger than their counterparts of need_a_name, even if p-values of the two methods are highly correlated (Pearson's correlation is 0.96). This is consistent with our earlier simulation (see results in Section 3.1.1) that CAMERA-modt could be too conservative. There is no systematic difference in p-values of GSEA and those of CAMERA-modt.

We report the top 30 enriched gene sets in Table 3. Five enriched gene sets identified by GSEA are also present (noted by "$*$" in the table). Originally, Labadorf et al. (2015) used the same HD data set to conduct enrichment analysis using topGo (Alexa and Rahnenfuhrer, 2010). They found that the enriched gene sets they identified show a clear immune response and inflammation-related pattern, including REACTOME INNATE IMMUNE SYSTEM, PID IL4 2PATHWAY, and

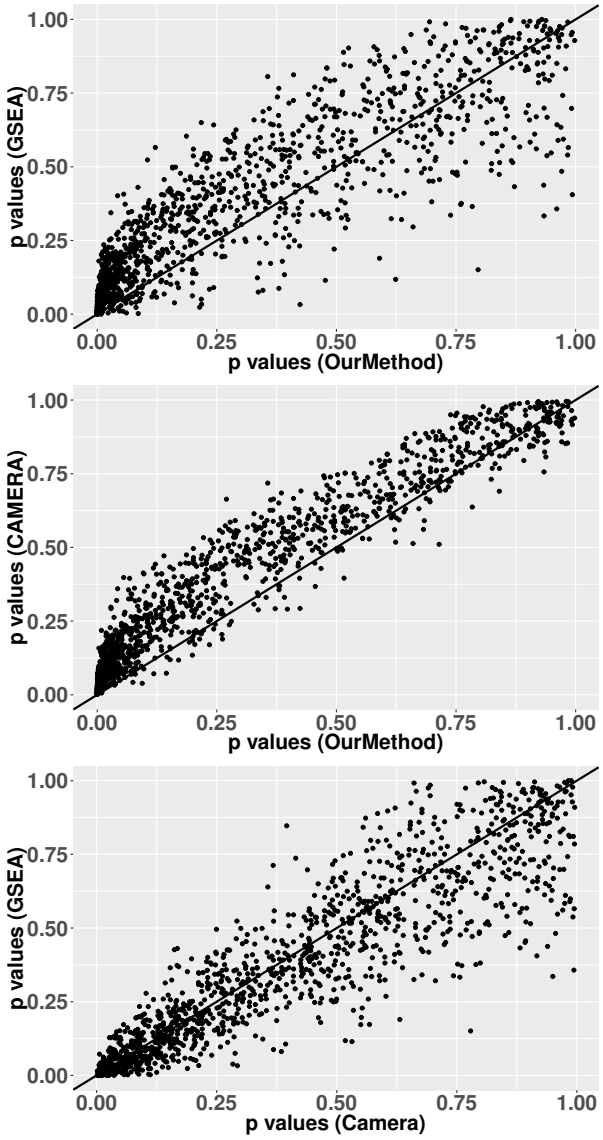

Figure 3: Pairwise comparisons of p-values for need_a_name, GSEA, and CAMERA-modt. The p-values are reported from enrichment test of each gene set in the C2 Canonical Pathway gene sets.

PID NFKAPPAB CANONICAL PATHWAY. These three gene sets rank 6,10 and 2 respectively in Table 3.

Many of our enriched gene sets have been shown to be closely related to HD pathogenesis. For example, the top enriched gene set by need_a_name, "PID SMAD2 3NUCLEAR PATHWAY" (see Table 3), is responsible for regulation of nuclear SMAD2/3 signaling. Katsuno et al. (2010) showed that nuclear SMAD2/3 are related to polyglutamine disease, which includes HD. The second enriched gene set, "PID NFKAPPAB CANONICAL PATHWAY", is a canonical NF-kappaB pathway, and its dysregulation causes HD immune dysfunction (Träger et al., 2014). Also, Marcora and Kennedy (2010) found that reduced transport of NF-kappaB out of dendritic spines and its activity in neuronal nuclei may contribute to the etiology of HD. Another gene set, "REACTOME INNATE IMMUNE SYSTEM", contributes to HD pathogenesis (Träger et al., 2014; Labadorf et al., 2015). Diamanti et al. (2013) showed that "REACTOME TRANSCRIPTIONAL ACTIVITY OF SMAD2 SMAD3 SMAD4 HETEROTRIMER", a gene set involved in transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer, is also enriched in their microarray study of HD pathology from blood samples of R6/2 at manifest stage and wild type littermate mice. For AKT signaling pathway, "BIOCARTA AKT PATHWAY", Humbert et al. (2002) demonstrated that huntingtin is a substrate of AKT and that phosphorylation of huntingtin by AKT is crucial to mediate the neuroprotective effects of IGF-1. They also showed that AKT is altered in Huntingtons disease patients. Chiang et al. (2010) demonstrated that the systematic downregulation of PPARγ, related to "BIOCARTA PPARA PATHWAY", seems to play a critical role in the dysregulation of energy homeostasis observed in HD, and that PPARγ is a potential therapeutic target for this disease. For "REACTOME SIGNALING BY TGF BETA RECEPTOR COMPLEX", Kandasamy et al. (2011) demonstrated that TGF-beta1 signaling appears to be a crucial modulator of neurogenesis in HD pathology and it can be a promising target for endogenous cell-based regenerative therapy. For "PID P53 DOWNSTREAM PATHWAY", Ghose et al. (2011) showed the likely involvement of NFkB (RelA), p53 and miRNAs in the regulation of cell death in HD pathogenesis.

3.2.2 *Male vs Female Lymphoblastoid Cells Data.* We analyzed the mRNA expression profiles from lymphoblastoid cell lines derived from 17 females and 15 males. Subramanian et al. (2005) examined this data set with their GSEA method, testing the enrichment of the cytogenetic gene sets (C1). C1 includes 24 sets, one for each of the 24 human chromosomes, and 295 sets corresponding to cytogenetic bands. For the comparison "male VS female", they expected to find gene sets on chromosome Y, not on chromosome X. We run enrichment analysis with three tests (need_a_name, GSEA and CAMERA-modt). In Table 4, we summarized all the gene sets with nominal p-value $\leqslant 0.01$ in at least one test. Three gene sets, one from chromosome Y and two Y bands, were found to be enriched by all three tests at FDR level 0.05. Interestingly, need_a_name identified another Y band, chrYp22, as enriched. In fact, the four gene sets called significant by need_a_name are the only four containing at least 3 genes in C1 and corresponding to chromosome Y or Y bands. need_a_name did not produce small p-value ($\leqslant 0.01$) for the remaining three gene sets in Table 4, which was just as expected in that study.

4. Conclusion and Discussion

(Conclusion) need_a_name is a mixed effects quasi-likelihood model for competitive gene set test. It effectively adjusts for completely unknown, unstructured correlations among the genes. It uses a score test approach and allows for analytical assessment of p-values. Compared to existing approaches, need_a_name controls type I error correctly and maintains good power under different correlation structures.

Table 3: Top 30 enriched gene sets using need_a_name for HD data. Gene sets are ranked by their associated p-values. FDR is the adjusted p-value using Benjamini-Hochberg (BH) procedure.

Gene Set	Size	ρ_1	ρ_2	ρ_3	p-value	FDR	
PID SMAD2 3NUCLEAR PATHWAY	79	0.071	0.011	0.017	7.5E-07	9.9E-04	$*$
PID NFKAPPAB CANONICAL PATHWAY	22	0.124	0.011	0.020	2.4E-06	1.6E-03	
REACTOME YAP1 AND WWTR1 TAZ STIMULATED GENE EXPRESSION	23	0.130	0.011	0.018	4.4E-06	1.7E-03	
REACTOME SIGNALING BY TGF BETA RECEPTOR COMPLEX	60	0.045	0.011	0.015	7.3E-06	1.7E-03	
BIOCARTA NTHI PATHWAY	23	0.124	0.011	0.024	7.5E-06	1.7E-03	
REACTOME INNATE IMMUNE SYSTEM	209	0.048	0.011	0.010	7.8E-06	1.7E-03	
KEGG PATHWAYS IN CANCER	311	0.029	0.011	0.010	8.9E-06	1.7E-03	
REACTOME DOWNSTREAM TCR SIGNALING	31	0.095	0.011	0.013	1.2E-05	1.9E-03	
KEGG NOD LIKE RECEPTOR SIGNALING PATHWAY	55	0.054	0.011	0.010	1.3E-05	1.9E-03	
PID IL4 2PATHWAY	59	0.086	0.011	0.012	1.4E-05	1.9E-03	
KEGG TGF BETA SIGNALING PATHWAY	82	0.062	0.011	0.013	2.7E-05	3.3E-03	
BIOCARTA 41BB PATHWAY	14	0.095	0.011	0.023	3.2E-05	3.4E-03	
PID P53 DOWNSTREAM PATHWAY	131	0.052	0.011	0.013	3.4E-05	3.4E-03	
REACTOME TCR SIGNALING	48	0.098	0.011	0.016	3.6E-05	3.5E-03	
REACTOME ACTIVATED TLR4 SIGNALLING	87	0.027	0.011	0.010	4.9E-05	4.2E-03	
REACTOME TOLL RECEPTOR CASCADES	110	0.038	0.011	0.010	5.2E-05	4.2E-03	
REACTOME TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION	69	0.015	0.011	0.010	5.4E-05	4.2E-03	
BIOCARTA TID PATHWAY	18	0.125	0.011	0.017	5.7E-05	4.2E-03	
BIOCARTA ALK PATHWAY	34	0.064	0.011	0.011	7.4E-05	5.1E-03	$*$
REACTOME SMAD2 SMAD3 SMAD4 HETEROTRIMER REGULATES TRANSCRIPTION	25	0.102	0.011	0.021	7.6E-05	5.1E-03	$*$
REACTOME TRANSCRIPTIONAL ACTIVITY OF SMAD2 SMAD3 SMAD4 HETEROTRIMER	36	0.079	0.011	0.021	8.3E-05	5.1E-03	
BIOCARTA AKT PATHWAY	20	0.023	0.011	0.010	8.8E-05	5.1E-03	$*$
ST TUMOR NECROSIS FACTOR PATHWAY	28	0.039	0.011	0.016	9.0E-05	5.1E-03	$*$
PID ANGIOPOIETIN RECEPTOR PATHWAY	50	0.082	0.011	0.013	9.3E-05	5.1E-03	
KEGG P53 SIGNALING PATHWAY	65	0.037	0.011	0.007	9.7E-05	5.1E-03	
KEGG APOPTOSIS	82	0.041	0.011	0.009	1.0E-04	5.1E-03	
BIOCARTA PPARA PATHWAY	53	0.026	0.011	0.008	1.1E-04	5.2E-03	
REACTOME MYD88 MAL CASCADE INITIATED ON PLASMA MEMBRANE	78	0.026	0.011	0.010	1.1E-04	5.2E-03	
PID BCR 5PATHWAY	64	0.064	0.011	0.016	1.2E-04	5.3E-03	
PID HIF1 TFPATHWAY	64	0.067	0.011	0.011	1.2E-04	5.3E-03	

ρ_1: average sample correlation between genes in the test set.

ρ_2: average sample correlation between genes in the background set.

ρ_3: average sample correlation between two genes, one from the test set and the other from the background set.

$*$: enriched gene sets identified by GSEA.

(What we proposed) Under competitive gene set test framework, a number of methods have been proposed to account for correlation among genes. One approach is to evaluate the set-level statistic by permuting sample labels to generate the null distribution, as adopted by the widely used GSEA (Subramanian et al., 2005). However, sample permutation method has been criticized for altering the null hypotheses being tested (Goeman and Bühlmann, 2007; Khatri et al., 2012). Instead, CAMERA (Wu and Smyth, 2012) proposed to correct for the correlation among genes by estimating a VIF directly from the data. This approach has also been used by Yaari et al. (2013) in their QuSAGE procedure. The major shortcoming with this approach is that it tries to estimate correlations among gene-level test statistics directly from sample correlation (is it clear??). In (the paper to be finished), we have argued that the correlations among gene-level statistics are not necessarily equal to those among samples due to the presence of DE genes. The estimated VIF could be biased without taking into account such a discrepancy and thus undermines the performance of CAMERA and QuSAGE. need_a_name avoids the discrepancy by using the differences in mean as gene-level statistics for a two group comparison experiment. It models the covariance of gene-level statistics by two variance components, one attributable to correlations among samples after treatment effect removed, and the other attributable to the DE effect associate with the treatment. We note that

Table 4: Summary of gene sets for lymphoblastoid cells data. Reported are gene sets with p-value $\leqslant 0.01$ for at least one of the need_a_name, GSEA, and CAMERA-modt methods. The FDR is the adjusted p-value using Benjamini-Hochberg (BH) procedure.

		need_a_name		GSEA		CAMERA-modt	
Gene set	Size	p-value	FDR	p-value	FDR	p-value	FDR
chrY	40	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.002
chrYq11	16	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
chrYp11	18	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.028
chrYp22	8	< 0.001	0.036	0.012	0.503	0.010	0.762
chr7p11	8	0.049	0.835	0.006	0.352	0.101	0.998
chr11p12	5	0.065	0.835	0.008	0.388	0.115	0.998
chrXp22	76	0.072	0.835	0.004	0.295	0.581	0.998

for need_a_name, the estimation of covariance among gene-level statistics need not be exact: need_a_name uses a score test that involves linear combinations of the entries in the covariance matrix. The denominator in the score test statistic (REF EQ) can usually be accurately approximated given the high dimensionality of the covariance matrix. need_a_name is based on quasi-likelihood, therefore it does not require normal assumption of expression data, and could be applied to both microarray and RNA-Seq experiments.

(Summarize the results) We compared need_a_name to other existing approaches in both simulation study and real data analysis. In the simulation study, we examined the performance of need_a_name and other six method (geneSetTest-modt, MRSGE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) in terms of type I error control and power. We demonstrated that under a variety of correlation structures, need_a_name holds correct type I error size and also maintains good power. In the real data analysis, need_a_name was able to identify more gene sets as enriched, some of which, in the corresponding studies, are insightful yet not found by methods such as GSEA or CAMERA.

(Future work) Currently, need_a_name only supports enrichment test for two-group comparisons. In many gene expression experiments, however, researchers might use more complex design to study different comparisons of interest, in which case a linear model would be more appropriate. Our future work will focus on generalizing need_a_name to allow for more complicated design structures.

The R codes for reproducing results in this paper are available at `https://github.com/zhuob/EnrichmentAnalysis`.

5. AcknowledgeMents

6. Appendix

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

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

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"),

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

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

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

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\leqslant j_1<j_2\leqslant n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22})\right] \\ &= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}) \\ &= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\ &\quad + \frac{n_2-1}{n_2}E\left(\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}|\Delta_i)\right) \\ &\quad + \frac{n_2-1}{n_2}\mathrm{Cov}\left(E(Y_{ij_12}|\Delta_i), E(Y_{ij_22}|\Delta_i)\right) \\ &= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)\end{aligned} \tag{15}$$

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

Alexa, A. and Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. *R Package Version*, 2(0).

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.

Chiang, M.-C., Chen, C.-M., Lee, M.-R., Chen, H.-W., Chen, H.-M., Wu, Y.-S., Hung, C.-H., Kang, J.-J., Chang, C.-P., Chang, C., et al. (2010). Modulation of energy deficiency in Huntington's disease via activation of the peroxisome proliferator-activated receptor gamma. *Human Molecular Genetics*, page ddq322.

Diamanti, D., Mori, E., Incarnato, D., Malusa, F., Fondelli, C., Magnoni, L., and Pollio, G. (2013). Whole gene expression profile in blood reveals multiple pathways deregulation in R6/2 mouse model. *Biomarker Research*, 1(1):1.

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.

Ghose, J., Sinha, M., Das, E., Jana, N. R., and Bhattacharyya, N. P. (2011). Regulation of miR-146a by RelA/NFkB and p53 in ST Hdh Q111/Hdh Q111 Cells, a Cell Model of Huntington's Disease. *PLoS One*, 6(8):e23837.

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.

Humbert, S., Bryson, E. A., Cordelières, F. P., Connors, N. C., Datta, S. R., Finkbeiner, S., Greenberg, M. E., and Saudou, F. (2002). The IGF-1/Akt pathway is neuroprotective in Huntington's disease and involves Huntingtin phosphorylation by Akt. *Developmental Cell*, 2(6):831–837.

Kandasamy, M., Reilmann, R., Winkler, J., Bogdahn, U., and Aigner, L. (2011). Transforming growth factor-beta signaling in the neural stem cell niche: a therapeutic target for Huntington's disease. *Neurology Research International*, 2011.

Katsuno, M., Adachi, H., Minamiyama, M., Waza, M., Doi, H., Kondo, N., Mizoguchi, H., Nitta, A., Yamada, K., Banno, H., et al. (2010). Disrupted transforming growth factor-β signaling in spinal and bulbar muscular atrophy. *The Journal of Neuroscience*, 30(16):5702–5712.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 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.

Labadorf, A., Hoss, A. G., Lagomarsino, V., Latourelle, J. C., Hadzi, T. C., Bregu, J., MacDonald, M. E., Gusella, J. F., Chen, J.-F., Akbarian, S., et al. (2015). RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *PloS One*, 10(12):e0143563.

Marcora, E. and Kennedy, M. B. (2010). The Huntington's disease mutation impairs Huntingtin's role in the transport of NF-κB from the synapse to the nucleus. *Human Molecular Genetics*, 19(22):4373–4384.

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. *Statistical Applications in Genetics and Molecular Biology*, 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.

Träger, U., Andre, R., Lahiri, N., Magnusson-Lind, A., Weiss,

A., Grueninger, S., McKinnon, C., Sirinathsinghji, E., Kahlon, S., Pfister, E. L., et al. (2014). HTT-lowering reverses Huntingtons disease immune dysfunction caused by NFκB pathway dysregulation. *Brain*, 137(3):819–833.

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.