Published online 31 July 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 12* ***1–13***
doi:10.1093/nar/gkn000

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

Bin Zhuo [1], Duo Jiang [2] *

[1,2]Department of Statistics, Oregon State University, 239 Weniger Hall, Corvallis, OR, 97333, USA

Received January 1, 20XX; Revised February 1, 20XX; Accepted March 1, 20XX

ABSTRACT

Competitive gene set test 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 inflated type I error rates and power loss. We propose a new framework, MEQLEA, for gene set test 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. Compared to existing methods such as 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 two real data analysis to illustrate our approach.

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 consisting 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 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 (Goeman and Bühlmann, 2007): the *self-contained* test and the *competitive* test . A self-contained test examines a set of genes by a fixed standard without reference to other genes in the genome (Goeman et al., 2005, 2004, Huang and Lin, 2013, Tsai and Chen, 2009, Wu et al., 2010). 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 convoluted 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 (Gatti et al., 2010, Goeman and Bühlmann, 2007).

Independent gene set test

Many competitive gene set tests rely on independence of gene-level statistics which further depends on independence among gene expression levels. 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, Gatti et al., 2010, Goeman and Bühlmann, 2007, Wu and Smyth, 2012, Yaari et al., 2013), as genes within a gene set are often co-expressed and function together.

*To whom correspondence should be addressed. Tel: +44 000 0000000; Fax: +44 000 0000000; Email: jiangd@oregonstate.edu

Tests that account for inter-gene correlation

A handful of methods have been proposed to account for inter-gene correlation in competitive gene set test. One attempt is to evaluate the set-level statistic by permuting the biological sample labels (Efron and Tibshirani, 2007, Subramanian et al., 2005). 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 inter-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 their set-level statistics to account for inter-gene correlations. QuSAGE (Yaari et al., 2013), which is a recent extension to CAMERA, also used the same VIF to adjust for inter-gene correlations but includes the VIF in a different set-level statistic. The VIF is a crucial factor and valid estimation of it relies on the assumption that correlation between any two gene-level statistics are almost the same as correlation between their corresponding expression levels. Barry et al. (2008) showed by simulation that this assumption holds 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, this assumption is likely to be problematic when a fraction of genes are truly DE, in which case the correlation among gene-level statistics (e.g., t-statistics) can be badly estimated by sample correlation (Zhuo and Di, unpublished work).

What do we propose?

We propose a new framework for enrichment analysis that we will call Mixed Effects Quasi-Likelihood Enrichment Analysis (MEQLEA). Our idea is motivated by the discrepancy between correlations among expression levels and those among gene-level statistics caused by the presence of DE genes. To tackle such discrepancy, 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. MEQLEA uses a score test approach and allows for analytical assessment of p-values. Compared to existing methods including GSEA and CAMERA, MEQLEA 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 we describe the methodology and then the simulation setup of MEQLEA, and summarize related existing methods; in Section we present results from comparison of MEQLEA to other existing methods by simulation study, and illustrate the application of our method by two real data sets; in Section we conclude and also specifies the future work.

METHODS

We consider a gene expression (e.g. RNA-Seq or microarray) experiment, in which we compare the gene expression levels of samples from two groups: a treatment group with n_1 samples referred to as "cases" and a control group with n_2 samples referred to as "controls" ($n_1, n_2 \geq 3$). Suppose the expression levels of a set of m genes are observed for each sample. An unknown subset of these genes are DE between cases and controls, with varying sign and magnitude of DE effects. The genes are also allowed to have (negatively or positively) correlated expression levels. In enrichment analysis, we are interested in a pre-defined set of genes, for example, from a known pathway or given by a functional annotation term from a database such as KEGG (Kanehisa and Goto, 2000) or GO (Ashburner et al., 2000). Our goal is to test whether this known gene set is enriched with differential expression signals. Let $\boldsymbol{G}$ be an m-dimensional vector defining the gene set of interest, where $G_i = 1$ if and only if the i^{th} gene is in the set and $G_i = 0$ otherwise. Our analysis will condition on $\boldsymbol{G}$ and test if $\boldsymbol{G}$ is associated with enhanced DE effects. In the following sections, we will first construct a hierarchical model for the gene expression data incorporating possible correlations among the m genes, from which we will derive a quasi-likelihood model for the gene-level DE statistics jointly for all the genes. Based on this model, we will then present our enrichment test, and discuss its connections with CAMERA. Finally, we will describe our simulation studies used to evaluate our method. For the rest of **Methods**, our presentation of the method is conditional on $\boldsymbol{G}$ unless otherwise indicated.

MEQLEA

A hierarchical model for the gene expression data We will start by presenting the hierarchical model for the observed gene expression data, which will incorporate the following features. Firstly, for a given sample, the expression levels of different genes are allowed to be correlated. We further assume that the correlation structure is the same across samples. Secondly, different genes may have different baseline expression levels, where "baseline" refers to the average among controls. Thirdly, for any given gene, its mean expression profile in the treatment group can be either higher, lower or the same compared to the control group, depending on whether the gene is up-regulated, down-regulated, or not DE. For the genes that are differentially expressed, their DE effects are modeled additively and are allowed to have heterogeneous signs and magnitudes. Finally, given a gene, and its DE effect, the expression profile is allowed to vary independently across samples, which captures measurement error and sample-level variability.

To present our model formally, we first introduce some notation. Let $n=n_1+n_2$ be the total sample size. Let $\boldsymbol{X}$ be an n-dimensional known vector of 1's and 0's denoting the case-control membership of the samples, with $X_i=1$ for a case and $X_i=0$ for a control. Let $\boldsymbol{Y}$ be an n by m matrix representing the expression data, in which each row is the expression profile for a sample and Y_{ij} $(1\leq i\leq n, 1\leq j\leq m)$ is the expression profile of sample i at gene j. Let μ_j $(1\leq j\leq m)$ be the baseline expression profile for gene j. μ_j's are treated as nuisance parameters and as we will see later do not contribute to our analysis. Let $\boldsymbol{\Delta}=(\Delta_1,\cdots,\Delta_m)^T$ be a vector for the additive DE effects for the genes. Gene j is not DE if $\Delta_j=0$, up-regulated if $\Delta_j>0$ and down-regulated if $\Delta_j<0$. We model $\boldsymbol{\Delta}$ as a random effect, for which we will detail our assumptions later. Given μ_j and Δ_j, the mean expression profile for the control group and the treatment group are μ_j and $\mu_j+\Delta_j$, respectively. Given these means, the noise in the observed expression data for the i^{th} sample is denoted by the mean zero error vector $\epsilon_i=(\epsilon_{i1},\cdots,\epsilon_{im})^T$, $1\leq i\leq n$. We assume $\boldsymbol{\epsilon}:=(\epsilon_1,\cdots,\epsilon_m)$ to be independent of $\boldsymbol{\Delta}$ and to have mean zero. Without loss of generality, we also assume $\text{Var}(\epsilon_{ij})=1$ for all genes and samples. For a real gene expression data set typically not satisfying this assumption, we can standardize the data by each gene to ensure its empirical variance equals one before implementing our method. For the covariance structure of ϵ, we assume

$$\epsilon_{i_1} \text{ and } \epsilon_{i_2} \text{ are independent}, \quad i_1\neq i_2, \tag{1}$$

$$\text{Cov}(\epsilon_i|\boldsymbol{G})=\boldsymbol{C},\ 1\leq i\leq n, \tag{2}$$

where $\boldsymbol{C}$ is an m by m inter-gene correlation matrix shared by all samples. $\boldsymbol{C}$ is generally unknown.

Putting these elements together, we obtain the following model for the expression data $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{G}$

$$Y_{ij}=\mu_j+X_i\cdot\Delta_j+\epsilon_{ij}, \tag{3}$$

for $1\leq i\leq n, 1\leq j\leq m$. $\boldsymbol{G}$ enters this model via Δ_j and possibly μ_j.

Assumptions on the DE effects Δ_j Conditional on $\boldsymbol{G}$, we assume that the Δ_j's are mutually independent and come from either of the two distributions, $\mathscr{D}_1$ and $\mathscr{D}_2$, depending on whether $G_j=0$ or 1. We denote the expected values of $\mathscr{D}_1$ and $\mathscr{D}_2$ by β_0 and $\beta_0+\beta_1$, respectively, and their variances by σ_1^2 and σ_2^2, respectively. It follows that

$$E(\boldsymbol{\Delta}|\boldsymbol{G})=\beta_0+\beta_1\boldsymbol{G},\ \text{var}(\boldsymbol{\Delta}|\boldsymbol{G})=\sigma_1^2\boldsymbol{I}_1+\sigma_2^2\boldsymbol{I}_2, \tag{4}$$

where $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are diagonal matrices of dimension m with 0's and 1's on their diagonals. The 1's in the diagonal of $\boldsymbol{I}_1$ correspond to the genes with $G_j=1$ and those for $\boldsymbol{I}_2$ to the genes with $G_j=0$.

Aside from the conditions in Equation (**4**) on the first two moments, we do not impose any specific distributional assumptions such as normality on $\boldsymbol{\Delta}$. For example, the distribution of a given Δ_j can put positive mass on zero, which allows for the highly likely event that some of the genes are not DE. To further motivate our general framework for $\boldsymbol{\Delta}$, we present a simple model included by Equation (**4**) as a special case. Suppose the m genes are independently sampled to be either DE or not. The probability for gene j to be DE is p_t if $G_j=1$ or p_b if $G_j=0$. For DE genes, their DE effects are sampled independently from a common distribution with mean μ_δ and variance σ_δ^2. Under these assumptions,

$$E(\Delta_j|\boldsymbol{G})=p_j\mu_\delta,\ \text{var}(\Delta_j|\boldsymbol{G})=p_j\sigma_\delta^2+p_j(1-p_j)\mu_\delta^2, \tag{5}$$

where $p_j=p_t$ if $G_j=1$ and $p_j=p_b$ if $G_j=0$. It can be shown that this model is a special case of Equation (**4**).

Model for gene-level statistics For each gene j, we consider the DE statistic (gene-level statistic???) given by

$$U_j=\frac{\sum_{i:X_i=1}Y_{ij}}{n_1}-\frac{\sum_{i:X_i=0}Y_{ij}}{n_2}, \tag{6}$$

which is sample mean difference in the expression profile between cases and controls. Given our assumption that ϵ_j has variance 1, U_j provides a DE metric for gene j. We will construct a quasi-likelihood model for $\boldsymbol{U}=(U_1,\cdots,U_m)^T$ by deriving the mean and covariance structures of $\boldsymbol{U}$ from the model for $\boldsymbol{Y}$ described in Sections and . We first observe that combining Equations (**6**) and (**3**) yields

$$U_j=\Delta_j+\eta_j, \text{where } \eta_j=\frac{1}{n_1}\sum_{i:X_i=1}\epsilon_{ij}-\frac{1}{n_2}\sum_{i:X_i=0}\epsilon_{ij}. \tag{7}$$

It can be shown based on Equations (**1**), (**2**) and (**4**) that

$$E(\boldsymbol{U}|\boldsymbol{G})=\beta_0+\beta_1\boldsymbol{G}, \tag{8}$$

$$\Sigma:=\text{Var}(\boldsymbol{U}|\boldsymbol{G})=\sigma_0^2\boldsymbol{C}+\sigma_1^2\boldsymbol{I}_1+\sigma_2^2\boldsymbol{I}_2, \tag{9}$$

where $\sigma_0^2=1/n_1+1/n_2$ is a known parameter. We note that in Equation (**9**), the covariance structure of $\boldsymbol{U}$ has three components, a component with $\boldsymbol{C}$ which accounts for the contribution from sample-level noise $\boldsymbol{\epsilon}$, and two additional components from the DE effect $\boldsymbol{\Delta}$. It is noteworthy that both the $\boldsymbol{C}$ component and the $\boldsymbol{\Delta}$ components contribute to the variance of U_j's, whereas only the $\boldsymbol{C}$ component contributes to the correlation among U_j's.

The set-level test statistic For a competitive gene set test, it is often unclear what the hypothesized null is and what is being tested (Barry et al., 2008, Wu and Smyth, 2012). In our approach, to detect patterns of the DE signals in the gene set of interest that stand out compared with genes not in the set, we test $H_0:\mathscr{D}_0=\mathscr{D}_1$ against $H_1:\mathscr{D}_0\neq\mathscr{D}_1$. For example, for the special scenario given by Equation (**5**), this amounts to testing $p_b=p_t$ against $p_b\neq p_t$. To construct the test statistic, we focus on the part of the alternative space where $E(\mathscr{D}_0)\neq E(\mathscr{D}_1)$, or equivalently $\beta_1\neq 0$. We first consider the less interesting case with uncorrelated genes, in which $\boldsymbol{C}$ equals $\boldsymbol{I}$, an m-dimensional identity matrix. Under the quasi-likelihood model for $\boldsymbol{U}$ given in Section , the quasi-score statistic for β_1 has the form $S\propto\boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\boldsymbol{1}_m)$, where $\hat{\beta}_0=\overline{U}$ is an estimate for β_0

and $\mathbf{1}_m$ is a m-dimensional vector of 1's. To perform a quasi-score test, one would divide S^2 by its estimated variance under H_0 and the assumption that $\boldsymbol{C}=\boldsymbol{I}$. The resulting test statistic is

$$T_{\text{u}}=\frac{S^2}{\widehat{\text{Var}}_{0,\boldsymbol{C}=\boldsymbol{I}}(S|\boldsymbol{G})}=\frac{[\boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}}, \quad (\mathbf{10})$$

where $\boldsymbol{H}=\frac{1}{n}\mathbf{1}_m\mathbf{1}_m^T$. The subscript "u" stands for "uncorrelated genes." For the case of interest when inter-gene correlation is present, $\boldsymbol{C}$ is a non-trivial correlation (???Covariance) matrix. We will again form our test statistic based on S. But for the denominator of the statistic, the null variance of S will be evaluated under the quasi-likelihood model with non-trivial $\boldsymbol{C}$. By Equation (**9**), the variance of S is given by $\text{Var}(S|\boldsymbol{G})=\boldsymbol{G}^T(\boldsymbol{I}-\boldsymbol{H})\Sigma(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}$. Note that $H_0:\mathscr{D}_0=\mathscr{D}_1$ implies $\sigma_1^2=\sigma_2^2$. Thus, under H_0, $\Sigma:=\text{Var}_0(\boldsymbol{U}|\boldsymbol{G})=\sigma_0^2\boldsymbol{C}+\sigma_1^2\boldsymbol{I}$, where $\sigma_0=1/n_1+1/n_2$ is known and σ_1^2 is an unknown parameter. To estimate σ_1^2 under H_0, we observe that $\text{var}_0(U_j)=\sigma_0^2+\sigma_1^2$ and use $\hat{\sigma}_1^2=\sum_{j=1}^m(U_j-\overline{U})^2/(m-1)-\sigma_0^2$. Therefore, assuming $\boldsymbol{C}$ is known, we can obtain the MEQLEA test statistic given by

$$T=\frac{S^2}{\widehat{\text{Var}}_0(S|\boldsymbol{G})}=\frac{[\boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I}-\boldsymbol{H})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}}, \quad (\mathbf{11})$$

where $\hat{\boldsymbol{\Sigma}}=(1/n_1+1/n_2)\boldsymbol{C}+\hat{\sigma}_1^2\boldsymbol{I}$ is a null estimate of $\boldsymbol{\Sigma}$. Under suitable regularity conditions, significance of the test could then be assessed by comparing T to a χ_1^2 distribution.

In practice, the inter-gene covariance matrix $\boldsymbol{C}$ is usually unknown. So we substitute $\boldsymbol{C}$ with $\hat{\boldsymbol{C}}$, the empirical covariance matrix of the expression data after controlling for possible DE effects by centering the expression levels of cases and controls separately around zero. Formally, $\hat{\boldsymbol{C}}$ is given by $\hat{C}_{jk}=\frac{1}{n}\sum_{i=1}^n(Y_{ij}-\alpha_{ij})(Y_{ik}-\alpha_{ik})$ where $\alpha_{ij}=\sum_{i':X_{i'}=X_i}Y_{i'j}/\sum_{i'=1}^n 1\{X_{i'}=X_i\}$ is the average expression profile at gene j for all samples from the same group (either treatment or control) as sample i. In real data sets, the number of genes, m, is usually much greater than the sample size n, in which case $\boldsymbol{C}$ is a high-dimensional parameter that cannot be efficiently estimated by $\hat{\boldsymbol{C}}$. Interestingly, however, we find (note???) that the the test statistic T relies not on the accurate estimation of the entire $\boldsymbol{C}$, but only on three parameters involving $\boldsymbol{C}$, which can be much more realistically estimated by a moderate sample size. To demonstrate this, we re-arrange the order of the rows and columns of $\boldsymbol{C}$ to allow the partition $\boldsymbol{C}=\begin{bmatrix}\boldsymbol{C}_{11} & \boldsymbol{C}_{12}\\ \boldsymbol{C}_{12}^T & \boldsymbol{C}_{22}\end{bmatrix}$, where $\boldsymbol{C}_{11}$ is the correlation matrix for genes in the test set, $\boldsymbol{C}_{22}$ is that for gene in the background set (i.e., the complement of the test set), and $\boldsymbol{C}_{12}$ is the cross-correlation matrix between the two classes of genes. (To be continued....)

Simulation Methods

Simulation Setup In this section, we will specify the parameter setup for type I error and power simulations.

Let $\boldsymbol{Y}_j$ be a vector denoting the expression profile of sample j and $\text{Cov}(Y_{i_1,j},Y_{i_2,j})=\rho_{i_1,i_2}$ for any two genes i_1 and i_2. We assume that genes have the same correlation if they are from the same category (whether the test set or the background set): $\text{Cov}(Y_{i_1},Y_{i_2})=\rho_1$ if genes i_1 and i_2 are both from the test set (i.e., $G_{i_1}=G_{i_2}=1$), $\text{Cov}(Y_{i_1},Y_{i_2})=\rho_2$ if they are both from the background set (i.e., $G_{i_1}=G_{i_2}=0$), and $\text{Cov}(Y_{i_1},Y_{i_2})=\rho_3$ if i_1 is from the test set and i_2 is from the background set (i.e., $G_{i_1}=1,G_{i_2}=0$). 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): $\rho_1=\rho_2=\rho_3=0.1$; that is, all genes are correlated, with an exchangeable correlation structure.

(c): $\rho_1=0.1$, $\rho_2=\rho_3=0$; that is, only the genes in the test set are correlated. This corresponds to ... , and we envision what methods do well...

(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.

By no means are such correlation structures intended to model the actual correlation structures among gene expression levels.

The simulations run as follows: first, we consider an entire gene set containing $m=500$ genes, of which $m_1=100$ genes are in the test set, and the remaining $m_2=400$ genes in the background set; second, we sample genes as DE with probability p_t in the test set and with probability p_b in the background set, and for sampled DE genes, we simulate the DE effect Δ from a normal distribution $N(2,1)$ (except in Table 2 we use $N(1,0.5)$ to report calibrated power) and for non-DE genes we set $\Delta=0$; 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 $\text{MVN}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ for the control group and n_2 samples from $\text{MVN}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ for the treatment group, where the covariance $\boldsymbol{\Sigma}=[\text{Cov}(Y_{i_1},Y_{i_2})]_{m\times m}$ may be one of the correlation structures in (a)-(e).

Further assumptions on p_t and p_b will complete our generating model used in the type I error and power simulations. (REF methods part about DE and no DE ....) 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. (a heads-up on how the 2 groups are different: if non-GO-term genes have DE) To illustrate this point, we perform 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 $0\%(S_0)$ in group A_1 and $10\%(S_0)$ 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 scenarios for the alternative hypothesis of the presence of enrichment: for genes in the test set, we set DE probability to be $5\%(S_1)$,$10\%(S_2)$,$15\%(S_3)$ and $20\%(S_4)$ in group A_1, and $15\%(S_1)$,$20\%(S_2)$,$25\%(S_3)$ and $30\%(S_4)$ in group A_2. Table 1 summarizes the simulation setup for the two groups.

Table 1. DE probability configurations in type I error and power simulations. S_0 is for type I error simulation. S_1-S_4 represent the four scenarios considered in power simulations. p_b and p_t are the DE probability for genes in the background set and that in the test set, respectively.

Group	Background DE prob. (p_b)	DE prob. in test set (p_t)				
		S_0	S_1	S_2	S_3	S_4
A_1	0%	0%	5%	10%	15%	20%
A_2	10%	10%	15%	20%	25%	30%

Other methods considered We will compare MEQLEA to six previously proposed gene set tests: GSEA (Subramanian et al., 2005), two versions of the CAMERA procedure —CAMERA-modt and CAMERA-rank (Wu and Smyth, 2012), SigPathway (Tian et al., 2005), MRGSE (Michaud et al., 2008), and QuSAGE (Yaari et al., 2013). Except SigPathway and MRGSE, all methods incorporate features intended for inter-gene correlation correction. GSEA calculates an enrichment score for the test set by examining the ranking (according to some metric, for example, signal-to-noise ratio) of its member genes, and determines the significance of the enrichment score by randomly permuting sample labels. CAMERA-modt uses moderated t-statistics (Smyth, 2004) as gene-level statistics and estimate a VIF to account for inter-gene correlations in the set-level statistic, and CAMERA-rank is the rank version of the CAMERA-modt. MRGSE is rank-based method assuming inter-gene independence, which is recommended by (Tarca et al., 2013) over a class of independence-assuming methods. SigPathway is a parametric version of MRGSE, and in this simulation we use the moderated t-statistics as the gene level statistics. QuSAGE generates from t-test a probability density function (PDF) for each gene, combines the individual PDFs using convolution, and quantifies enrichment of the test set with a complete PDF.

For software implementation, GSEA is modified from the original R-GSEA script (http://software.broadinstitute.org/gsea/index.jsp) to accommodate single gene set test. CAMERA and MRGSE are implemented in the limma package (Smyth, 2005) in the Bioconductor project (Gentleman et al., 2004), QuSAGE is available in the Bioconductor package of the same name, and SigPathway is implemented by ourselves. (Move to discussion)Because GSEA and MEQLEA do not support linear models, the implementations are restricted to two-group comparisons.

In terms of type I error control and power, we expect some of the six tests to have different performances between group A_1 and A_2 simulations under one or more correlation structures.

RESULTS

According to the simulation setup in Section , the test set is not enriched if DE probabilities are the same for 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), in which case we evaluate the type I error. As to power, we set DE probability according to each of the alternative scenarios S_1-S_4 (see Table 1) and calculate the proportion of data sets for which a test would reject at a given level α. The results are based on $10{,}000$ simulated data sets.

Type I error simulations

Messages: 1. Our method is well calibrated for all scenarios. 2. All of these others methods have poor calibration in at least some of the scenarios. 3. The way type I error is deviates for the nominal level aligns with our expectations for GSEA (permutation-based), CAMERA (rank-based or parametric T-test), and methods assuming independence (GRSGE and SigPath)

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 (MEQLEA, SigPathway, MRGSE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) under each of the five correlation structures (each row of plots, from top to bottom, corresponds accordingly to correlation structures (a)-(e)).

In group A_1 simulations (the left column of Figure 1), GSEA and MEQLEA 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 versions of CAMERA control type I errors correctly for correlation structures (a) and (c), yet they are too conservative for the case of (b) and anti-conservative for correlation structures (d) and (e). SigPathway and MRGSE procedures have well-calibrated type I error for correlation structures (a) and (b), but are anti-conservative the case of (c), (d) and (e). QuSAGE has good type I error control for only (c), and is too conservative for (a), (d) and (e), and anti-conservative for (b).

In group A_2 simulations (the right column of Figure 1), MEQLEA continues to hold the size of type I error rate, whereas GSEA is 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, but may be liberal in other situations SigPathway and MRGSE have similar trends for p-values as they do, respectively, in group A_1 simulations. QuSAGE is conservative in (b) but anti-conservative in the remaining four correlation structures.

Explain why this happens

MEQLEA 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 inter-gene correlation structures, and DE probability of each gene. MEQLEA controls the size of type I error well because it uses difference in mean as gene-level statistic, and the correlations

between such statistics are exactly the same as correlations between the samples (Zhuo and Di, unpublished work). 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 simulations, 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 inter-gene correlation structure, which explains the complete failure of GSEA in controlling type I error for the case of group A_2 simulations. For CAMERA and QuSAGE, 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 (Zhuo and Di, unpublished work), and therefore the set-level test statistic is under-estimated. The performances of related methods—QuSAGE and two versions of CAMERA—are subject to the underlying correlation structures. (Considering remove this sentence...) Moreover, the performance of CAMERA is complicated by the fact that the set-level statistic takes into account only the inter-gene correlation in the test set without addressing that in the background set.

Different from the five methods mentioned above, SigPathway and MRGSE rely on independence between genes. It's not surprising that such gene permutation based methods control type I error correctly when genes are "equally-correlated": in (a) genes are simulated to be independent, and in (b) genes are simulated to have an exchangeable correlation structure. However, both SigPathway and MRGSE fail to hold type I error size for the remaining three correlation structures. These simulations show that even small inter-gene correlations will result in inflated type I error rate when the test does not account for inter-gene correlations.

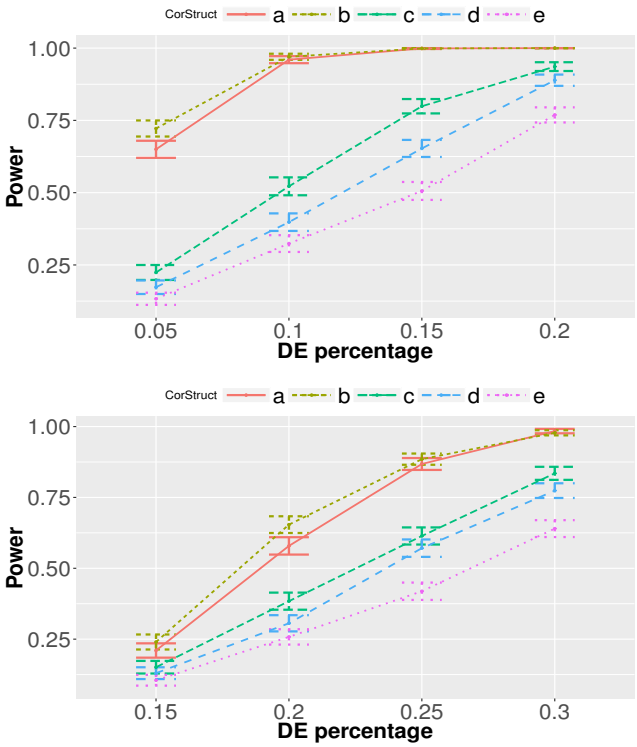

Figure 2. Power for MEQLEA under correlation structures (a)-(e) of Section . The top corresponds to group A_1 simulations, and the bottom to group A_2 simulations (see Table 1). The error bars are the 95% CIs based on 10,000 simulations.

Power simulation

Messages: 1. Correlation structure among genes does impact power (Fig. 2). 2. If the genes are independent (not likely...), our method has better or similar power than independence-assuming methods. GSEA is more powerful when there is no background DE, but fails miserably otherwise.
What puzzles us: In Fig. 2, structure b (rho=0.1 for all genes) yields higher power than structure a (uncorrelated genes).

We compare the power of MEQLEA to those of the other six methods under correlation structure (a) in which genes are simulated to be independent. Since some of these tests are not well calibrated at the sample size considered (see results in 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 2 summarizes the calibrated power for the two groups of simulations (i.e., A_1 and A_2 in Table 1). For A_1 simulations, GSEA has the highest, and rank based methods (MRGSE and CAMERA-rank) have the lowest, calibrated power across all four alternative scenarios. CAMERA-modt, SigPathway and MEQLEA have no systematic difference in the calibrated power. In group A_2 simulations, GSEA shows virtually no power. MEQLEA, CAMERA-modt, and SigPathway have indistinguishable calibrated power and are among the best.

Figure 2 shows for MEQLEA, 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 under correlation structures (a) and (b) 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 (b) are simulated to have the same correlation 0.1. As the correlation structure becomes more complex, from (c) 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.

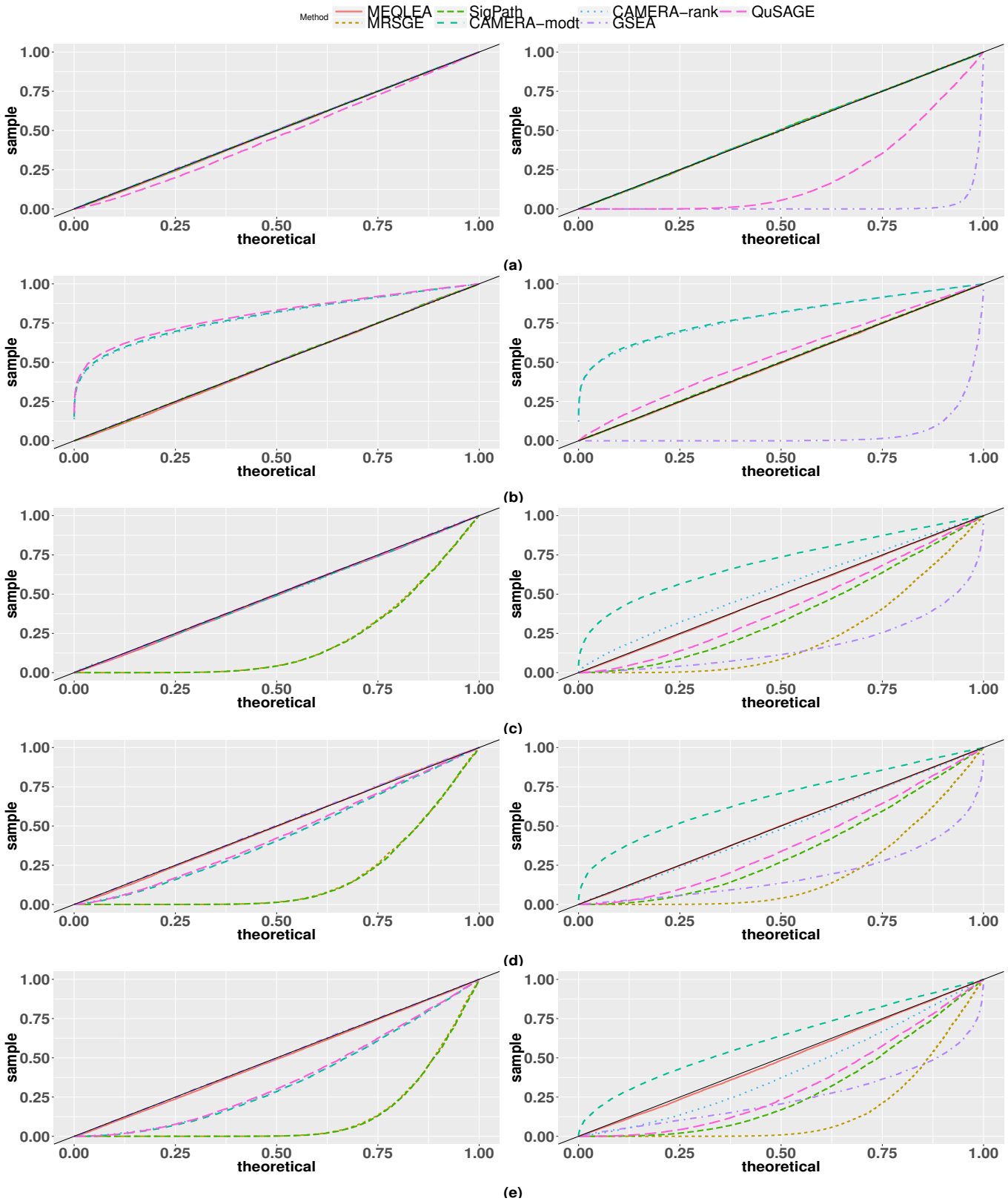

Figure 1. Uniform quantile-quantile plots for p-values by different methods. Each plot from top to bottom corresponds to correlation structures (a)-(e), respectively. The left column is for group A_1 simulation, and the right column for group A_2 simulation (see Table 1 for detail). Results are based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

Table 2. Recalibrated power (standard error) for different methods. The powers are summarized under four alternatives S_1-S_4 in each of the group A_1 and A_2 simulations (see Table 1 for detail). Results are based on $10{,}000$ simulations.

Group	Method	$c(\alpha)$	S_1	S_2	S_3	S_4
A_1	MEQLEA	0.045	0.340	0.741	0.944	0.991
	MRGSE	0.051	0.111	0.284	0.533	0.766
	SigPathway	0.049	0.344	0.744	0.947	0.992
	CAMERA-modt	0.051	0.336	0.737	0.943	0.990
	CAMERA-rank	0.053	0.108	0.280	0.519	0.758
	GSEA	0.051	0.517	0.894	0.989	0.999
	QuSAGE	0.028	0.385	0.784	0.959	0.995
A_2	MEQLEA	0.050	0.180	0.478	0.777	0.939
	MRGSE	0.048	0.104	0.269	0.530	0.781
	SigPathway	0.049	0.175	0.473	0.773	0.936
	CAMERA-modt	0.052	0.173	0.466	0.766	0.933
	CAMERA-rank	0.050	0.102	0.262	0.521	0.771
	GSEA	0.000	0.000	0.000	0.000	0.000
	QuSAGE	0.000	0.021	0.127	0.387	0.692

Real Data

We applied MEQLEA to two example data sets, and compared the lists of enriched gene sets to those obtained by other three methods (GSEA, CAMERA-modt and MRGSE). Our results lend credence to previous studies in finding potential gene sets correlated with Huntington's disease and those correlated with chromosome Y and Y bands in lymphoblastoid cells.

Huntington's Disease Data **Summary: 1. Our method yields more signals than CAMERA and GSEA, but fewer than MRGSE. 2. P-values by our method are generally less conservative than those by CAMERA. For genes with small p-values, our method tends to be conservative than MRGSE, and less so than GSEA. 3. The top 30 gene sets identified by our method contain some plausible biologically functions linked to Huntington's disease.**

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/). For each gene, we adjusted for two covariates–age at death (DeathAge) and RNA Integrity Number (RIN), as also done by Labadorf et al. (2015). We followed their strategy of treating the two covariates as categorical. Briefly, DeathAge was binned into intervals 0-45, 46-60, 61-75, 76-90 and 90+, and RIN was dichotomized as $>$ or ≤ 7. We regressed the normalized expression levels on AgeDeath and RIN and use the resulting residuals as the *covariate-adjusted expression levels*.

We performed enrichment analysis on the covariate-adjusted data 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.

We applied four test procedures (MEQLEA, GSEA, CAMERA-modt and MRGSE) to run enrichment analysis for the entire C2 Canonical Pathway gene sets, and compared the four tests in terms of resulting enriched gene sets. We used the Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedure (BH) to control the false discovery rate (FDR) for multiple hypothesis testing (unless specified otherwise, all p-values in Section were adjusted by BH procedure). The BH procedure is used when the test statistics under the null have non-negative correlations (Benjamini and Yekutieli, 2001). We note that since many pathways have overlapped genes, the BH procedure should be appropriate in our study.

In Figure 3 we plot $\log 10$ p-values for MEQLEA against GSEA, CAMERA-modt and MRGSE. The p-values of CAMERA-modt are overwhelmingly larger than their counterparts of GSEA or of MEQLEA, yet smaller than those of MRGSE, even if p-values between MEQLEAand other three methods are highly correlated (Pearson's correlation of $\log 10$ p between MEQLEA and GSEA, CAMERA-modt and MRGSE are 0.90, 0.96, and 0.87 respectively). (Can we say this? we don't know the truth) This is consistent with our earlier simulation (see results in simulation section) that CAMERA-modt could be too conservative. The p-values of MRGSE are in general larger than the corresponding p-values of MEQLEA.

Using MEQLEA, we found 89 significant signals out of the entire 1330 gene sets at FDR level of 0.05. GSEA found 3 enriched gene sets—2 of them were also among those 89 gene sets (the one that was not significant according to MEQLEA had a p-value of 0.013 and FDR 0.100). MRGSE found 387 gene sets which include all the 89

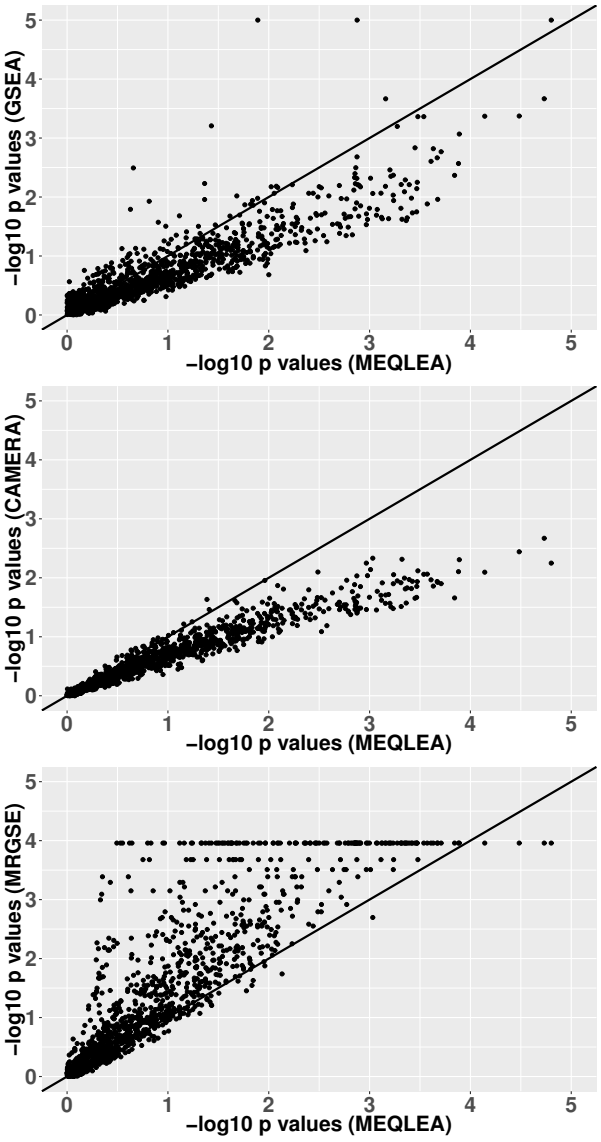

Figure 3. Pairwise comparisons of p-values for MEQLEA, GSEA, and CAMERA-modt. The p-values are reported from enrichment test of each gene set in the C2 Canonical Pathway gene sets.

sets MEQLEA identified, and CAMERA-modt found none. 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 "PID NFKAPPAB CANONICAL PATHWAY". These three gene sets rank 18,10 and 3 (by nominal p-values) respectively in Table 3.

In Table 3, we report the top 30 enriched gene sets (ordered by nominal p values) identified using MEQLEA. We also labeled the enriched gene sets from GSEA by "$*$" in the table. Many of our enriched gene sets have been shown to be closely related to HD pathogenesis. For example, the top enriched gene set by MEQLEA,"PID SMAD2 3NUCLEAR PATHWAY", 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 third 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 (Labadorf et al., 2015, Träger et al., 2014). 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 "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.

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 the four tests (MEQLEA, GSEA,CAMERA-modtand MRGSE). In Table 4, we summarized all the gene sets that are called significant at FDR level 0.05. Unanimously, three gene sets—"chrY", "chrYq11" and "chrYp11"—were found to be enriched by all of the four methods. Interestingly, only MEQLEA was able to identify another Y band, "chrYp22", as enriched. In fact, these four gene sets are the only four pathways containing at least 3 genes in C1 and corresponding to chromosome Y or Y bands. MEQLEA did not produce small p-value (<0.01) for the remaining three gene sets in Table 4, which was just as expected in that study.

CONCLUSION AND DISCUSSION

(Conclusion) MEQLEA 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, MEQLEA controls type I error correctly and maintains good power under different correlation structures.

(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,

Table 3. Enriched gene sets (ordered by nomial p-values) identified by MEQLEA for HD data. The $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_3$, respectively, are the average estimated sample correlation between genes in the test set, between genes in the background set, and between two genes–one from the test set and the other from the background set. The enriched gene sets are noted by "$*$" for GSEA. No gene set was identified as enriched by CAMERA-modt and all the 30 gene sets are also identified as enriched by MRGSE. For all methods, a gene set is called significant when its FDR using Benjamini-Hochberg (BH) correction is < 0.05.

Gene Set	Size	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	p-value	FDR	
PID SMAD2 3NUCLEAR PATHWAY	79	0.063	0.013	0.015	5.8E-06	5.7E-03	$*$
REACTOME YAP1 AND WWTR1 TAZ STIMULATED GENE EXPRESSION	23	0.121	0.013	0.014	8.5E-06	5.7E-03	
PID NFKAPPAB CANONICAL PATHWAY	22	0.127	0.013	0.019	2.3E-05	1.0E-02	
BIOCARTA NTHI PATHWAY	23	0.130	0.013	0.023	6.2E-05	2.1E-02	
BIOCARTA TID PATHWAY	18	0.101	0.013	0.012	1.2E-04	2.2E-02	
PID HIV NEF PATHWAY	35	0.065	0.013	0.013	1.2E-04	2.2E-02	
KEGG PATHWAYS IN CANCER	311	0.028	0.013	0.010	1.3E-04	2.2E-02	
PID MYC REPRESS PATHWAY	60	0.057	0.013	0.013	1.9E-04	2.2E-02	
BIOCARTA TOLL PATHWAY	36	0.083	0.013	0.018	2.0E-04	2.2E-02	
PID IL4 2PATHWAY	59	0.081	0.013	0.010	2.0E-04	2.2E-02	
KEGG TGF BETA SIGNALING PATHWAY	82	0.055	0.013	0.011	2.2E-04	2.2E-02	
BIOCARTA DEATH PATHWAY	33	0.067	0.013	0.013	2.4E-04	2.2E-02	
KEGG NOD LIKE RECEPTOR SIGNALING PATHWAY	55	0.045	0.013	0.008	2.6E-04	2.2E-02	
BIOCARTA CTCF PATHWAY	23	0.083	0.013	0.015	2.8E-04	2.2E-02	
ST TUMOR NECROSIS FACTOR PATHWAY	28	0.031	0.013	0.014	3.2E-04	2.2E-02	
BIOCARTA TNFR2 PATHWAY	17	0.151	0.013	0.022	3.3E-04	2.2E-02	
KEGG APOPTOSIS	82	0.036	0.013	0.008	3.3E-04	2.2E-02	
REACTOME INNATE IMMUNE SYSTEM	209	0.039	0.013	0.009	3.3E-04	2.2E-02	
PID HES HEY PATHWAY	47	0.071	0.013	0.019	3.4E-04	2.2E-02	
REACTOME DOWNSTREAM TCR SIGNALING	31	0.082	0.013	0.011	3.7E-04	2.2E-02	
PID TCPTP PATHWAY	42	0.076	0.013	0.010	3.7E-04	2.2E-02	
BIOCARTA 41BB PATHWAY	14	0.110	0.013	0.023	3.9E-04	2.2E-02	
PID FRA PATHWAY	34	0.154	0.013	0.008	4.1E-04	2.2E-02	
PID P53 DOWNSTREAM PATHWAY	131	0.045	0.013	0.012	4.2E-04	2.2E-02	
PID EPO PATHWAY	34	0.069	0.013	0.013	4.3E-04	2.2E-02	
BIOCARTA PPARA PATHWAY	53	0.031	0.013	0.008	4.4E-04	2.2E-02	
BIOCARTA EPONFKB PATHWAY	11	0.068	0.013	0.010	4.7E-04	2.2E-02	
BIOCARTA HIVNEF PATHWAY	58	0.063	0.013	0.019	4.8E-04	2.2E-02	
BIOCARTA CD40 PATHWAY	13	0.165	0.013	0.026	4.8E-04	2.2E-02	
BIOCARTA IL7 PATHWAY	17	0.100	0.013	0.016	5.2E-04	2.3E-02	

Table 4. Enriched gene sets and their nominal p values for lymphoblastoid cells data. Reported are gene sets with FDR< 0.05 for at least one of the MEQLEA, GSEA, CAMERA-modt and MRGSE methods using Benjamini-Hochberg(BH) procedure.

Gene set	Size	MEQLEA	GSEA	CAMERA-modt	MRGSE
chrY	40	0.0E+00	0.0E+00	1.0E-05	5.9E-07
chrYq11	16	0.0E+00	0.0E+00	7.2E-08	8.5E-06
chrYp11	18	2.1E-15	0.0E+00	2.8E-04	5.1E-04
chrYp22	8	3.6E-04	1.2E-02	1.0E-02	1.3E-02
chr6	614	5.6E-02	6.0E-01	6.1E-01	2.1E-04
chr1	1104	6.1E-02	5.5E-01	6.3E-01	5.3E-05
chr12	571	8.7E-02	2.6E-01	4.0E-01	5.1E-09

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??). Zhuo and Di have argued (unpublished work) 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. MEQLEA 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 for MEQLEA, the estimation of covariance among gene-level statistics need not be exact: MEQLEA 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. MEQLEA 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 MEQLEA to other existing approaches in both simulation study and real data analysis. In the simulation study, we examined the performance of MEQLEA and other six method (SigPathway, MRGSE, 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, MEQLEA holds correct type I error size and also maintains good power. In the real data analysis, MEQLEA 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, MEQLEA 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 MEQLEA to allow for more complicated design structures.

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

ACKNOWLEDGMENTS

We thank Yanming Di, Sarah Emerson and Wanli Zhang for helpful discussion.

Conflict of interest statement. None declared.

APPENDIX

Standardization Standardization for each gene: first, we obtain the residuals by subtracting off the means within each treatment group;

$$r_{ijk} = y_{ijk} - \sum_{j=1}^{n_k} y_{ijk}/n_k; \tag{12}$$

then we calculate the pooled standard deviation from the residuals,

$$s_i = std(r_{ijk}); \tag{13}$$

next we get the standardized expression by dividing the original expression levels by the standard deviation,

$$y^*_{ijk} = y_{ijk}/s_i \tag{14}$$

We perform the standardization procedure to every gene in the data set.

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

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

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"),

$$\begin{aligned}\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})] \\ &\quad + \mathrm{Cov}[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})] \\ &= E\left(\frac{1}{n_1}\rho_{i_1,i_2} + \frac{1}{n_2}\rho_{i_1,i_2}\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}\end{aligned} \tag{15}$$

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

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

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22})\right] \\ &= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}) \\ &= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\ &+ \frac{n_2-1}{n_2}E\left(\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}|\Delta_i)\right) \\ &+ \frac{n_2-1}{n_2}\mathrm{Cov}\left(E(Y_{ij_12}|\Delta_i), E(Y_{ij_22}|\Delta_i)\right) \\ &= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)\end{aligned} \tag{16}$$

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

$$\mathrm{Cov}(\boldsymbol{T}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{17}$$

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).

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

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.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

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.

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.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

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.

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*

case	MEQLEA	MRGSE	SigPathway	CAMERA-modt	CAMERA-rank	GSEA	QuSAGE
a0PCT	0.056	0.049	0.051	0.049	0.047	0.049	0.078
a10PCT	0.050	0.052	0.051	0.048	0.050	0.946	0.491
b0PCT	0.059	0.050	0.051	0.000	0.000	0.048	0.000
b10PCT	0.052	0.051	0.051	0.000	0.000	0.837	0.027
c0PCT	0.056	0.513	0.517	0.051	0.044	0.051	0.052
c10PCT	0.054	0.442	0.188	0.000	0.021	0.290	0.131
d0PCT	0.059	0.586	0.594	0.114	0.104	0.051	0.106
d10PCT	0.052	0.522	0.235	0.001	0.049	0.220	0.175
e0PCT	0.058	0.674	0.679	0.213	0.197	0.053	0.203
e10PCT	0.054	0.614	0.334	0.004	0.116	0.113	0.267

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. *PloS one*, 8(11):e79217.

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.