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, 97331, USA

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

ABSTRACT

Competitive gene-set analysis is a widely used tool for interpreting high-throughput biological data, such as gene expression and proteomics data. It aims at testing a known category 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 mis-calibrated type I error rates and/or power loss. We propose a new framework, MEACA, 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 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 I error and maintains good power in a variety of correlation structure and association settings. We also present two real data analyses to illustrate our approach.

INTRODUCTION

Gene-set tests (also called enrichment analysis in some literature) aim to evaluate the association between the expression levels of genes in a pre-defined set, referred to as the test set, and experimental or environmental factors of interest. It examines whether the test set is enriched (or depleted) with differential expression (DE) signals, where the DE signal of a gene can be quantified by comparing the gene's expression levels across treatment groups for the factor of interest. The test set could be a known pathway or given by a functional annotation term from a database such as KEGG (20) or GO (21). Gene-set tests help researchers understand the underlying biological processes in terms of ensembles of genes.

Depending on the definition of the null hypothesis, there are two types of gene-set tests (1): the *self-contained* test and the *competitive* test. A self-contained test examines the DE signals of genes in the test set without reference to other genes in the genome (2, 3, 4, 5, 6). A competitive test compares DE signals of genes in the test set to those of genes not in the test set (7, 8, 9). Many methods, regardless of the type of test, perform a three-stage analysis (10): at 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 statistics include, among others, *signal-to-noise ratio* (11), *ordinary t-statistic* (7) or *moderated t-statistic* (12), *log fold change* (13) and *Z-score* (14). At 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 statistics are *enrichment score* (11), *maxmean statistic* (15), and a statistic derived from convoluted distribution of gene-level statistics (9), to name a few. At 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 literature (1, 16).

Many competitive gene-set tests rely on independence between gene-level statistics, which implicitly requires independence among the expression level of different genes. Examples of independence-assuming gene-set tests include, among many others, PAGE (13), the contingency-table-based tests (see **(author?)** (17) for a review) and sigPathway(7, 12). However, between-gene correlations can be widespread, for example, among co-regulated genes (16). Even mild correlations may result in inflated false positive rate for independence-assuming gene-set tests (1, 8, 9, 15, 16).

A handful of methods have been proposed to account for between-gene correlations in competitive gene-set tests. One attempt is to evaluate the set-level statistic by permuting the biological sample labels (11, 15). Permuting sample labels does not require an explicit understanding of the underlying correlation structure among genes and thus protects the test against such correlations. Since permuting sample labels is computationally inefficient, **(author?)** (18) 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, permuting sample labels in these methods inevitably alters the null and alternative hypotheses being tested, and therefore confuses the competitive test with the self-contained test, making the results hard to interpret (1, 8, 10). Another attempt is to use set-level statistics that directly include between-gene correlations estimated from the data. For example, CAMERA (8) calculates a *variance inflation factor* (VIF) from sample correlations (after the treatment effects removed) of observed data, and then incorporates it into their set-level statistics to account for between-gene correlations. QuSAGE (9), which is a recent extension to CAMERA but quantifies gene-set activity with a probability density function, also uses a similar

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

VIF to handle between-gene correlations. However, the VIF approach may be problematic in two respects: first and most importantly, it does not properly model the heterogeneity among genes in terms of the presence and magnitude of DE effects. Second, the VIF quantifies the strength of correlation among test statistics (e.g., t-statistics), but is approximated from sample correlation of observed data. Such approximation assumes that test-statistic correlations are almost the same as sample correlations of observed data, which might be violated when a fraction of genes are truly differentially expressed (Zhuo, Jiang and Di, unpublished work). We will show that the VIF approach can lead to severely compromised type I error and power in gene-set testing.

We propose a new framework for competitive gene-set test that we will call "**M**ixed-effects **E**nrichment **A**nalysis with **C**orrelation **A**ccounted for" (abbreviated to "MEACA"). Our idea is motivated by the discrepancy between correlations among expression levels and those among gene-level statistics caused by the presence of differentially expressed genes (Zhuo, Jiang and Di, unpublished work). We model the covariance structure of gene-level statistics by two variance components, one attributable to correlations among samples after treatment effect removed, and the other attributable to the variability across genes in the DE effects associated with the treatment. We use a quasi-likelihood model, which does not require the distribution of gene expressions or the distribution of the DE effects across genes to be Gaussian. Our method effectively adjusts for completely unknown, unstructured correlations among the genes. MEACA uses a score test approach and allows for analytical assessment of p-values. Compared to existing methods including GSEA (11) and CAMERA (8), MEACA enjoys robust and improved control over type I error and maintains good power in a variety of correlation structure and association settings.

The rest of the paper is organized as follows: in Section we describe the methodology of MEACA, as well as the simulation setup for evaluating type I error rate and power, and then we summarize some existing methods; in Section we present simulation results to compare MEACA to other methods, and illustrate the use of our method by two real data sets; in Section we conclude and also specify some future work.

METHODS

We consider a gene expression (e.g. RNA-Seq or microarray) experiment, in which we compare the 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 differentially expressed 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 (20) or GO (21). Our goal is to test whether this known gene set is enriched with DE signals. We will refer to the pre-defined gene set as "the test set," and genes not in this set " the background genes" which make up "the background set." We will use a gene-level test statistic, denoted by U_i, to capture the unknown DE signal of gene i. 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 test set and $G_i = 0$ otherwise (for any given gene set $\boldsymbol{G}$ is known). As an overview, in the following sections, we will derive a model for U_i's conditional on $\boldsymbol{G}$, using a mixed-effects framework of the form (details to be explained later)

$$U_i = \beta_0 + \beta_1 \mathrm{G}_i + \psi_i + \eta_i,\ i = 1, \ldots, m, \qquad (\mathbf{1})$$

where β_1 is a fixed effect capturing the mean difference between the test set and the background set, and ψ_i and η_i are random effects. The term ψ_i captures the variability among U_i's due to some genes being differentially expressed and some not, and to the varying magnitude of the DE effects. The variance of ψ_i depends on $G_i = 0$ or 1, which allows the spread of gene-level statistics to be different between the test set and the background set. The η_i's account for the variability in U_i's due to the across-sample variability, and are allowed to be correlated to incorporate between-gene covariations.

To justify model (**1**) and to elaborate the modeling assumptions on ψ_i and η_i, we will start by constructing a hierarchical model for the observed gene expression data, from which we will derive the mixed-effects model (**1**) for the gene-level 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 this section, our presentation of the method is conditional on $\boldsymbol{G}$ unless otherwise indicated.

MEACA

A hierarchical model for 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 level 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 differentially expressed. 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 level 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 m by n matrix representing the expression data, in which each column is the expression profile for a sample and Y_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$ is the expression level of sample j at gene i. Let μ_i $(1 \leq i \leq m)$

be the baseline expression level for gene i. The quantities μ_i'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 i is not differentially expressed if $\Delta_i=0$, up-regulated if $\Delta_i>0$ and down-regulated if $\Delta_i<0$. We model $\boldsymbol{\Delta}$ as a random effect, for which we will detail our assumptions later. Given μ_i and Δ_i, the mean expression level for the control group and the treatment group are μ_i and $\mu_i+\Delta_i$, respectively. Given these means, the noise in the observed expression data for the j^{th} sample is denoted by the mean zero error vector $\epsilon_j=(\epsilon_{1j},\cdots,\epsilon_{mj})^T$, $1\leq j\leq n$. 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 that its empirical variance equals one before implementing our method (see Appendix () for more detail). For the covariance structure of $\boldsymbol{\epsilon}$, we assume

$$\epsilon_{j_1} \text{ and } \epsilon_{j_2} \text{ are independent}, \quad j_1\neq j_2, \tag{2}$$

$$\text{Cov}(\epsilon_j|\boldsymbol{G})=\boldsymbol{C},\ 1\leq i\leq n, \tag{3}$$

where $\boldsymbol{C}$ is an m by m between-gene correlation matrix shared by all samples and is generally unknown. Putting these elements together, we obtain the following model for the expression data $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{G}$

$$Y_{ij}=\mu_i+X_j\cdot\Delta_i+\epsilon_{ij}, \tag{4}$$

for $1\leq i\leq m, 1\leq j\leq n$. The term $\boldsymbol{G}$ enters this model via Δ_i and possibly μ_i.

Assumptions on the DE effects Conditional on $\boldsymbol{G}$, we assume that the Δ_i's are mutually independent and come from either of the two distributions, $\mathscr{D}_1$ for statistics in the test set (i.e, $G_i=1$) and $\mathscr{D}_2$ for statistics in the background set (i.e, $G_i=0$). 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

$$E(\boldsymbol{\Delta}|\boldsymbol{G})=\beta_0+\beta_1\boldsymbol{G},\ \text{var}(\boldsymbol{\Delta}|\boldsymbol{G})=\sigma_1^2\boldsymbol{I}_1+\sigma_2^2\boldsymbol{I}_2, \tag{5}$$

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 $G_i=1$ and those for $\boldsymbol{I}_2$ to the genes with $G_i=0$.

Aside from the conditions in equation (**5**) on the first two moments, we do not impose any specific distributional assumptions such as normality on the DE effects $\boldsymbol{\Delta}$. For example, the distribution of a given Δ_i can put positive mass on zero, which allows for the highly likely event that some of the genes are not differentially expressed. To further motivate our general framework for $\boldsymbol{\Delta}$, we present a simple model included by equation (**5**) as a special case. Suppose the m genes are independently sampled to be either differentially expressed or not. The probability for gene i to be differentially expressed is p_t if $G_i=1$, or p_b if $G_i=0$. For differentially expressed genes, their DE effects are sampled independently from a common distribution with mean μ_δ and variance σ_δ^2. Under these assumptions,

$$E(\Delta_i|\boldsymbol{G})=p_i\mu_\delta,\ \text{Var}(\Delta_i|\boldsymbol{G})=p_i\sigma_\delta^2+p_i(1-p_i)\mu_\delta^2, \tag{6}$$

where $p_i=p_t$ if $G_i=1$ and $p_i=p_b$ if $G_i=0$ (the derivation is provided in Appendix ()). It can be shown that this model is a special case of equation (**5**).

Model for gene-level statistics For each gene i, we consider the gene-level statistic U_i given by

$$U_i=\frac{\sum_{j:X_j=1}Y_{ij}}{n_1}-\frac{\sum_{j:X_j=0}Y_{ij}}{n_2}, \tag{7}$$

which is sample mean difference in the expression levels between cases and controls. Given our assumption that ϵ_i has variance 1, U_i provides a DE metric for gene i. 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 (**7**) and (**4**) yields

$$U_i=\Delta_i+\eta_i,\text{where } \eta_i=\frac{1}{n_1}\sum_{j:X_j=1}\epsilon_{ij}-\frac{1}{n_2}\sum_{j:X_j=0}\epsilon_{ij}. \tag{8}$$

It can be shown based on equations (**2**), (**3**) and (**5**) that

$$E(\boldsymbol{U}|\boldsymbol{G})=\beta_0+\beta_1\boldsymbol{G}, \tag{9}$$

$$\Sigma:=\text{Var}(\boldsymbol{U}|\boldsymbol{G})=\sigma_0^2\boldsymbol{C}+\sigma_1^2\boldsymbol{I}_1+\sigma_2^2\boldsymbol{I}_2, \tag{10}$$

where $\sigma_0^2=1/n_1+1/n_2$ is a known parameter (the proof is provided in Appendix ()). We note that in equation (**10**), 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 effects $\boldsymbol{\Delta}$. It is noteworthy that both the $\boldsymbol{C}$ component and the $\boldsymbol{\Delta}$ components contribute to the variance of U_i's, whereas only the $\boldsymbol{C}$ component contributes to the correlations among U_i's.

We note that by letting $\Delta_i=\beta_0+\beta_1G_i+\psi_i$, equation (**8**) is equivalent to model (**1**) whose mean and variance are given by equations (**9**) and (**10**). The random effects ψ_i's capture the heterogeneity of the DE effects that are conditional on whether gene i belongs to the test set ($G_i=1$) or not ($G_i=0$).

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 (8, 19). 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 $H_0:\mathscr{D}_1=\mathscr{D}_2$ against $H_1:\mathscr{D}_1\neq\mathscr{D}_2$. For example, for the special scenario given by equation (**6**), this amounts to testing $p_b=p_t$ against $p_b\neq p_t$. To construct the set-level test statistic, we focus on the part of the alternative space where $E(\mathscr{D}_1)\neq E(\mathscr{D}_2)$, 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 $S \propto \boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\mathbf{1}_m)$, 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

$$T_{\text{u}}=\frac{S^2}{\widehat{\text{Var}}_{0,\boldsymbol{C}=\boldsymbol{I}}(S|\boldsymbol{G})}=\frac{[\boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}}, \qquad \textbf{(11)}$$

where $\boldsymbol{H}=\frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$. The subscript "u" stands for "uncorrelated genes". For the case of interest when between-gene correlation is present, $\boldsymbol{C}$ is a non-trivial correlation matrix. We will again form our test statistic based on S. However, 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 (**10**), 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 $H_0:\mathscr{D}_1=\mathscr{D}_2$ 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 $\sigma_0^2=1/n_1+1/n_2$ is known and σ_1^2 is an unknown parameter. To estimate σ_1^2 under H_0, we observe that $\text{Var}_0(U_i)=\sigma_0^2+\sigma_1^2$ and use $\hat{\sigma}_1^2=\sum_{i=1}^m(U_i-\overline{U})^2/(m-1)-\sigma_0^2$. Therefore, assuming $\boldsymbol{C}$ is known, we can obtain the MEACA test statistic given by

$$T=\frac{S^2}{\widehat{\text{Var}}_0(S|\boldsymbol{G})}=\frac{[\boldsymbol{G}^T(\boldsymbol{U}-\hat{\beta}_0\mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I}-\boldsymbol{H})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}}, \qquad \textbf{(12)}$$

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 between-gene covariance matrix $\boldsymbol{C}$ is usually unknown. Therefore 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 $\hat{C}_{ik}=\frac{1}{n}\sum_{j=1}^n(Y_{ij}-\alpha_{ij})(Y_{kj}-\alpha_{kj})$ where $\alpha_{ij}=\sum_{j':X_{j'}=X_j}Y_{ij'}/\sum_{j'=1}^n 1\{X_{j'}=X_j\}$ is the average expression level at gene i for all samples from the same group (either treatment or control) as sample j. 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 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 given a moderate sample size. To demonstrate this, let m_1 and m_2 be the size of the tested set and the background set, respectively ($m_1+m_2=m$). Let also ρ_1 be the average correlation between two genes in the tested set, ρ_2 the average correlation between two background genes, and ρ_3 the average correlation between a tested gene and a background gene. Then, ρ_1 is the mean of the off-diagonal entries in the $m_1\times m_1$ sub-matrix of $\boldsymbol{C}$ made up of rows and columns corresponding to the test set, ρ_2 is that in the $m_2\times m_2$ sub-matrix corresponding to the background set, and ρ_3 is the mean of the entries in the $m_1\times m_2$ sub-matrix of $\boldsymbol{C}$ corresponding to the cross-covariance between the tested and the background sets. It can be shown that the denominator of the MEACA test statistic given in equation (**12**) can be written as

$$c_1(1-\frac{1}{m_1})\rho_1+c_1(1-\frac{1}{m_2})\rho_2-2c_1\rho_3+c_1(\frac{1}{m_1}+\frac{1}{m_2})+c_2\hat{\sigma}_1^2, \qquad \textbf{(13)}$$

where $c_1=\frac{m_1^2m_2^2}{m^2}(\frac{1}{n_1}+\frac{1}{n_2})$ and $c_2=\frac{m_1m_2}{m}$. Therefore, the MEACA test statistic depends on $\boldsymbol{C}$ only through ρ_1, ρ_2 and ρ_3.

Connection with CAMERA Model (**1**) and equation (**13**) also help reveal the connections between CAMERA and our method (we use CAMERA-modt to illustrate). The test statistic for CAMERA-modt can be viewed as a score test derived from our model with the DE random effect ϕ_i omitted. Under our framework, CAMERA-modt effectively assumes that

1. Between-gene correlations are present only among genes in the test set, which means $\rho_2=\rho_3=0$;
2. $\psi_i=0$ for both genes in the test set and those in the background set, which amounts to assuming that $\sigma_1^2=\sigma_2^2=0$.

When the CAMERA approach is considered with $\text{Var}(\epsilon_{ij})=1$ and equation (**7**) as the gene-level statistics, its set-level test statistic can be shown to be equivalent to equation (**12**) but using the following as the denominator

$$c_1(1-\frac{1}{m_1})\rho_1+c_1(\frac{1}{m_1}+\frac{1}{m_2}). \qquad \textbf{(14)}$$

As a results, CAMERA does not properly model the contribution of between-gene correlations outside the test set, as well as non-zero, heterogeneous DE effects on the variance of the distribution of gene-level statistics.

To gain insights into the type I error performance of CAMERA-modt, we consider the following hypothetical scenarios where only the first assumption of CAMERA-modt is violated, that is, we allow at least one of ρ_2 and ρ_3 to be nonzero. Note that when m_1 and m_2 are large, the first term in equation (**13**) is approximately proportional to $\rho_1+\rho_2-2\rho_3$. We claim that (i) if $\rho_2>0$ but $\rho_3=0$, CAMERA-modt would under-estimate the denominator (because $\rho_1+\rho_2-2\rho_3>\rho_1$) and therefore would have inflated type I error ; (ii) if $\rho_2=\rho_3>0$, CAMERA-modt would be too conservative in controlling type I error ($\rho_1+\rho_2-2\rho_3<\rho_1$); (iii) if $\rho_2>0$ but $\rho_3<0$, CAMERA-modt would also produce inflated type I error ($\rho_1+\rho_2-2\rho_3>\rho_1$). These trends will be confirmed and illustrated using type I error simulations (see Section). On the other hand, if there are DE genes in both the test set and the background set (this corresponds to violation of the second assumption), the test statistic of CAMERA-modt is not comparable to that of MEACA.

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 $\mathrm{Cov}(Y_{i_1,j},Y_{i_2,j})=\rho_{i_1,i_2}$ 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): $\mathrm{Cov}(Y_{i_1},Y_{i_2})=\rho_1$ if genes i_1 and i_2 are both from the test set (i.e., $G_{i_1}=G_{i_2}=1$), $\mathrm{Cov}(Y_{i_1},Y_{i_2})=\rho_2$ if they are both from the background set (i.e., $G_{i_1}=G_{i_2}=0$). For cross-category genes, $\mathrm{Cov}(Y_{i_1},Y_{i_2})=\rho_3$ if i_1 is from the test set and i_2 is from the background set (i.e., $G_{i_1}=1,G_{i_2}=0$). 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.

(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 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 to be differentially expressed with probability p_t in the test set and with probability p_b in the background set, and for sampled differentially expressed genes, we simulate the DE effects $\boldsymbol{\Delta}$ from a normal distribution $N(2,1)$ (except in Table 3 we use $N(1,0.5)$ to report calibrated power) and for non-differentially expressed genes we set $\Delta_i=0$; third, we set the "true" mean expression values $\boldsymbol{\mu}_1=\boldsymbol{0}_m$ 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).

Further assumptions on p_t and p_b will complete our generating model used in the type I error and power simulations. We have mentioned that the test-statistic correlations may not be approximated by their sample correlations of observed data when at least one gene is truly differentially expressed (Zhuo, Jiang and Di, unpublished work). Therefore, if there are differentially expressed genes in the entire genome, approaches assuming almost equality between test-statistic correlations and sample correlations of observed-data may not perform well. For each of (a)-(e) correlation structures, we conduct two groups of simulations: genes in the background set are allowed to be differentially expressed in group A_2 but not in group A_1. 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 MEACA to six previously proposed gene-set tests: GSEA (11), two versions of the CAMERA (8) procedure—CAMERA-modt and CAMERA-rank, SigPathway (7), MRGSE (22), and QuSAGE (9). Except SigPathway and MRGSE, all methods incorporate features intended for between-gene correlation correction. GSEA calculates an enrichment score for the test set by examining the ranking (according to some metric, for example, the 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 (12) as gene-level statistics and estimates a VIF to account for between-gene correlations in the set-level statistic, and CAMERA-rank is the rank version of the CAMERA-modt. MRGSE is a rank-based method assuming between-gene independence, and is recommended by **(author?)** (23) 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 by the convoluted PDF that has incorporated the VIF to account for between-gene correlations.

The software implementation is described as follows. The 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 (24) in the Bioconductor project (25), and SigPathway is implemented by ourselves—these three methods use moderated-t as gene-level statistics. QuSAGE is available in the Bioconductor package of the same name and we use the default options.

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 examine the type I error rate. 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

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 (MEACA, 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)). We also summarize the type I error rates of all the methods at a significant level of 0.05 in Table 2.

In group A_1 simulations (the left column of Figure 1), GSEA and MEACA 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 for 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), MEACA 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.

MEACA 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. GSEA evaluates the enrichment score of a test set by generating its null distribution from sample permutation. When there are no differentially expressed 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 differentially expressed 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 simulations. As we've mentioned in the Introduction part, the VIF approach does not properly model the heterogeneity among genes in terms of the presence and magnitude of DE effects, and the VIF may be over-estimated (for moderated t-statistic in CAMERA) when a fraction of genes are differentially expressed (Zhuo, Jiang and Di, unpublished work). Therefore, the performances of VIF related methods—QuSAGE and two versions of CAMERA—are subject to the underlying DE effects.

Different from the five methods mentioned above, SigPathway and MRGSE rely on independence between genes. It's not surprising that such gene-label 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 between-gene correlations (e.g., 0.05) will result in inflated type I error rate when the test does not account for between-gene correlations.

Power simulation

We compare the power of MEACA 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 3 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 MEACA have no systematic difference in the calibrated power. In group A_2 simulations, GSEA shows virtually no power. MEACA, CAMERA-modt, and SigPathway have indistinguishable calibrated power and are among the best.

Figure 2 shows for MEACA, 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.

Real Data

We apply MEACA to two example data sets, and compare the lists of enriched gene sets to those obtained by other three

Table 2. Type I error rates for different methods under correlation structures (a)—(e). The type I error rates are summarized at a significant level of 0.05, based on $10,000$ simulated data sets.

Group	Correlation	MEACA	MRGSE	SigPathway	CAMERA-modt	CAMERA-rank	GSEA	QuSAGE
A_1	(a)	0.056	0.049	0.051	0.049	0.047	0.049	0.078
	(b)	0.059	0.050	0.051	0.000	0.000	0.048	0.000
	(c)	0.056	0.513	0.517	0.051	0.044	0.051	0.052
	(d)	0.059	0.586	0.594	0.114	0.104	0.051	0.106
	(e)	0.058	0.674	0.679	0.213	0.197	0.053	0.203
A_2	(a)	0.050	0.052	0.051	0.048	0.050	0.946	0.491
	(b)	0.052	0.051	0.051	0.000	0.000	0.837	0.027
	(c)	0.054	0.442	0.188	0.000	0.021	0.290	0.131
	(d)	0.052	0.522	0.235	0.001	0.049	0.220	0.175
	(e)	0.054	0.614	0.334	0.004	0.116	0.113	0.267

Table 3. Recalibrated power for different methods under correlation structure (a) in which genes are simulated to be independent. The $c(\alpha)$ is the α quantile of null distribution of p-values. The powers are calculated as the proportion of $p < c(\alpha)$ for each of the four alternatives S_1-S_4 and 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	MEACA	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

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 We examine the Huntington's Disease (HD) RNA-Sequencing (RNA-Seq) data (26) to identify enriched gene sets that are potentially responsible for HD. The mRNA expression profiles in human prefrontal cortex were obtained from 20 Huntington's Disease samples and 49 neurologically normal controls. Expression values are normalized and filtered as described in the methods section of **(author?)** (26). 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 adjust for two covariates—age at death (DeathAge) and RNA Integrity Number (RIN), as also done by **(author?)** (26). We follow their strategy of treating the two covariates as categorical. Briefly, DeathAge is binned into intervals 0-45, 46-60, 61-75, 76-90 and 90+, and RIN is dichotomized as $>$ (coded as 1) or ≤ 7 (coded as 0). We regress the normalized expression levels on AgeDeath and RIN and use the resulting residuals as the *covariate-adjusted expression levels*.

We perform enrichment analysis on the covariate-adjusted data using the MsigDB (11) C2 Canonical Pathways (February 5, 2016, data last accessed). The C2 Canonical Pathways 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 are named by HGNC symbols in C2 and by ensembl IDs in the HD expression data set, we convert the ensembl IDs in the expression data into HGNC symbols using *BioMart* (http://uswest.ensembl.org/biomart/martview/). We retain $26,941$ genes that have corresponding HGNC symbols.

We apply four test procedures (MEACA, GSEA, CAMERA-modt and MRGSE) to run enrichment analysis for the entire C2 Canonical Pathways, and compared the four tests in terms of resulting enriched gene sets. We use the Benjamini-Hochberg (27) procedure (BH) to control the false discovery rate (FDR) for multiple hypothesis testing (unless specified otherwise, all p-values in Section are adjusted by BH procedure). The BH procedure is used when the test

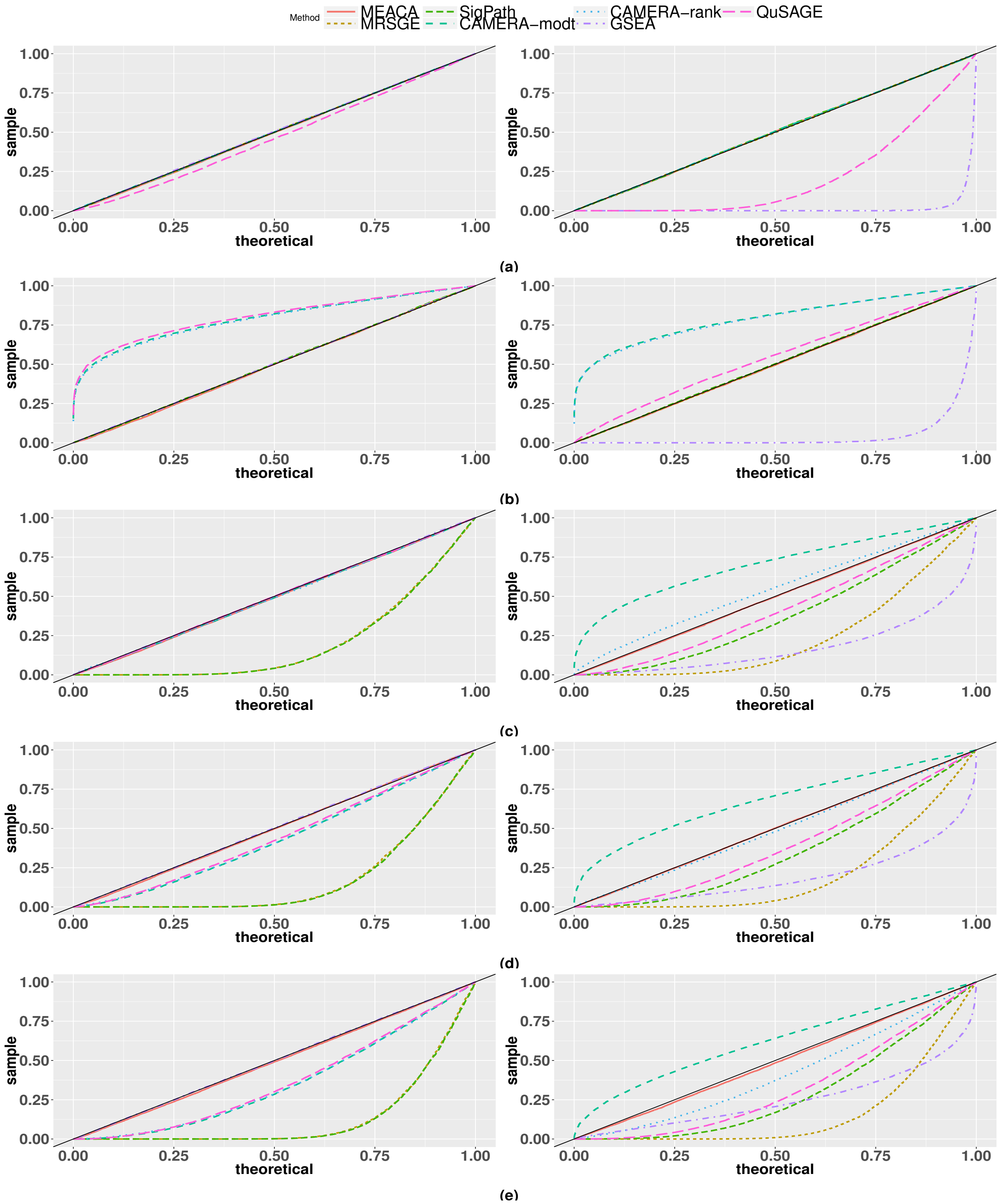

Figure 1. Uniform quantile-quantile plots for p-values by different methods. Each plot from top to bottom corresponds to correlation structures (a)-(e), respectively. The left column is for group A_1 simulation, and the right column for group A_2 simulation (see Table 1 for detail). Results are based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

statistics under the null have non-negative correlations (28).

We note that since many pathways have overlapped genes, the BH procedure should be appropriate in our study.

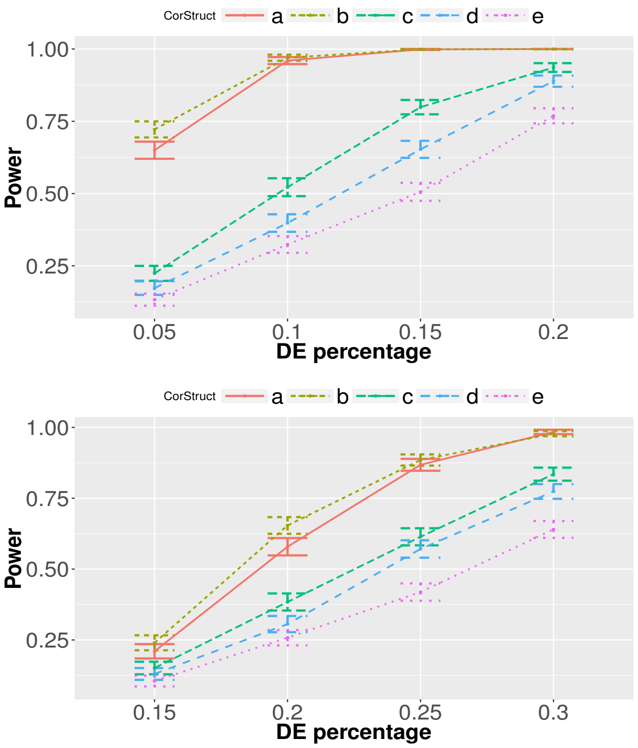

Figure 2. Power for MEACA under correlation structures (a)-(e) of Section . The top corresponds to group A_1 simulations, and the bottom to group A_2 simulations (see Table 1). The error bars are the 95% CIs based on 10,000 simulations.

In Figure 3 we plot $-\log 10$ p-values of MEACA against those of GSEA, CAMERA-modt and MRGSE. The p-values of CAMERA-modt are overwhelmingly larger than their counterparts of GSEA or MEACA, yet smaller than those of MRGSE, even if p-values between MEACA and other three methods are highly correlated (Pearson's correlation of $\log 10$ p between MEACA and GSEA, CAMERA-modt and MRGSE are 0.90, 0.96, and 0.87 respectively). This trend of p-values is consistent with our earlier simulation (see results in simulation section) that CAMERA-modt could produce large p values. The p-values of MRGSE are in general smaller than the corresponding p-values of MEACA, leading to more significant calls.

Using MEACA, we find 89 significant signals out of the entire 1330 gene sets at FDR level of 0.05. GSEA finds 3 enriched gene sets—2 of them were also among those 89 gene sets (the one that is not significant according to MEACA had a p-value of 0.013 and FDR 0.100). MRGSE identified 387 gene sets which include all the 89 sets MEACA identified, and CAMERA-modt identified none. Originally, **(author?)** (26) used the same HD data set to conduct enrichment analysis using topGo (29). They note 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 (by nominal p-values) 18, 10 and 3 respectively in the 89 enriched gene sets.

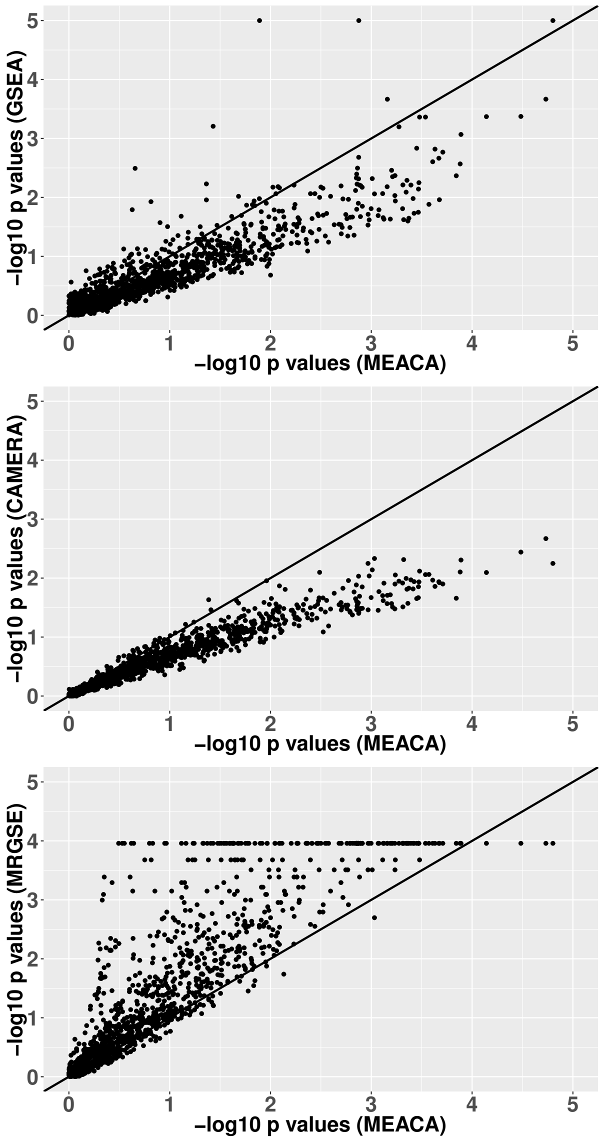

Figure 3. Pairwise comparisons of p-values for MEACA, GSEA, CAMERA-modtand MRGSE. The p-values are reported from enrichment test of each gene set in the C2 Canonical Pathway gene sets.

In Table 4, we report the top 30 enriched gene sets (ordered by nominal p values) identified using MEACA. We also label 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 MEACA, "PID SMAD2 3NUCLEAR PATHWAY," is responsible for regulation of nuclear SMAD2/3 signaling. **(author?)** (30) 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 (31). Also, **(author?)** (32) 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 (26, 31). **(author?)** (33) 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," **(author?)** (34) 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 analyze the mRNA expression profiles from lymphoblastoid cell lines derived from 17 females and 15 males. **(author?)** (11) examined this data set with their GSEA method, testing the enrichment of the cytogenetic gene sets (C1). The 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 (MEACA, GSEA,CAMERA-modtand MRGSE). In Table 5, we summarize all the gene sets that are called significant at FDR level 0.05 by at least one of the four test procedures. Unanimously, three gene sets—"chrY," "chrYq11" and "chrYp11"—are found to be enriched by all of the four methods. It is interesting to note that only MEACA is 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. MEACA does not produce small p-value (< 0.01) for the remaining three gene sets in Table 5, which is just as expected in that study.

CONCLUSION AND DISCUSSION

MEACA 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, MEACA controls type I error correctly and maintains good power under different correlation structures.

Under the 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 procedure GSEA (11). However, sample permutation method has been criticized for altering the null hypotheses being tested (1, 10). Instead, CAMERA (8) proposed to correct for the correlations among genes by estimating a VIF directly from the data. Incorporating VIF into set-level test statistics has also been used by **(author?)** (9) in their QuSAGE procedure where they quantify the gene set activity by a probability density function. The problems with this VIF approach are that it does not properly model the heterogeneity among genes in terms of the presence and magnitude of DE effect, and that it is intended to account for test-statistic correlations but is estimated from sample correlations of observed data. We have argued (in Chapter **??**) that the correlations among gene-level statistics may not be approximated by the sample correlation of observed data due to the presence of differentially expressed genes. Such problems will undermine the performances of CAMERA and QuSAGE. In contrast, MEACA 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 heterogeneity of DE effects. We note that for MEACA, the estimation of covariance among gene-level statistics need not be exact: MEACA uses a score test that involves linear combinations of the entries of the covariance matrix. The denominator in the score test statistic (see equation (**12**)) can usually be accurately approximated given the high dimensionality of the covariance matrix. MEACA 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.

We compare the performance of MEACA to those of other existing approaches through both simulation study and real data analyses. In the simulation study, we examine the calibration of MEACA and other six method (SigPathway, MRGSE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) in terms of type I error control and power. We demonstrate that MEACA holds correct type I error size under all correlation structures considered, whereas all other methods may fail in one or more situations. MEACA is also among the best in terms of power performance even under independent correlation structure. In the real data analyses, we run enrichment analysis using four methods—MEACA, CAMERA-modt, GSEA and MRGSE—on two data sets. The p-values of MEACA are smaller than those produced by GSEA and CAMERA that are intended to adjust for between-gene correlations, but larger than those of MRGSE that assumes independence between genes. MEACA is able to identify a moderate size of enriched gene sets, many of which are confirmed by independent studies yet are not revealed by other three methods.

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

The work of BZ in this article was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977 (to YD, SCE, and JHC). We thank Yanming Di, Sarah Emerson and Wanli Zhang for helpful discussion in preparing this manuscript. We thank Dr. Adam Labadorf for providing information about the HD gene expression data.

Conflict of interest statement. None declared.

Table 4. Enriched gene sets (ordered by nomial p-values) identified by MEACA for HD data. The $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_3$, respectively, are the average estimated sample correlations of observed data between genes in the test set, between genes in the background set, and between cross-category genes. 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 5. 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 MEACA, GSEA, CAMERA-modt and MRGSE methods using Benjamini-Hochberg (BH) procedure.

Gene set	Size	MEACA	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

APPENDIX

Standardization

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

$$r_{ijk}=y_{ijk}-\sum_{j=1}^{n_k} y_{ijk}/n_k; \qquad \textbf{(15)}$$

then we calculate the pooled standard deviation from the residuals,

$$s_i=std(r_{ijk}); \qquad \textbf{(16)}$$

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

$$y^*_{ijk} = y_{ijk}/s_i \tag{17}$$

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

Covariance matrix for test statistics

Note that if we let δ_i be the DE size for gene i, then $E(\delta_i) = \mu_\delta$ and $\mathrm{Var}(\delta_i) = \sigma^2_\delta$. To prove the equation (**6**), we introduce an additional random variable Z_i for DE status, where $Z_i \sim \mathrm{Bernoulli}(1,p_t)$ if $G_i = 1$ and $Z_i \sim \mathrm{Bernoulli}(1,p_b)$ if $G_i = 0$. It follows that $\Delta_i = Z_i\delta_i$, and we have

$$E(\Delta_i|\boldsymbol{G}) = E(Z_i\delta_i|\boldsymbol{G}) = E(\delta_i)E(Z_i|\boldsymbol{G}) = p_i\mu_\delta,$$

and

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

If the gene-level test statistics U_i's take the form of equation (**7**), then we have $E(U_i|\boldsymbol{G}) = E(\Delta_i + \eta_i|\boldsymbol{G}) = p_i\mu_\delta$. Next, note that the covariance between two genes i_1 and i_2 is given by

$$\begin{aligned}\mathrm{Cov}(U_{i_1},U_{i_2}|\boldsymbol{G}) &= \mathrm{Cov}[(\Delta_{i_1}+\eta_{i_1},\Delta_{i_2}+\eta_{i_2})|\boldsymbol{G}]\\ &= \mathrm{Cov}(\Delta_{i_1},\Delta_{i_2}|\boldsymbol{G}) + \mathrm{Cov}(\eta_{i_1},\eta_{i_2}|\boldsymbol{G})\\ &= \mathrm{Cov}\Bigg[\Big(\frac{1}{n_1}\sum_{j:X_j=1}\epsilon_{i_1,j} - \frac{1}{n_2}\sum_{j:X_j=0}\epsilon_{i_1,j},\\ &\qquad \frac{1}{n_1}\sum_{j:X_j=1}\epsilon_{i_2,j} - \frac{1}{n_2}\sum_{j:X_j=0}\epsilon_{i_2,j}\Big)|\boldsymbol{G}\Bigg]\\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)c_{i_1,i_2},\end{aligned} \tag{18}$$

where c_{i_1,i_2} is the corresponding entry in $\boldsymbol{C}$. Also

$$\begin{aligned}\mathrm{Var}(U_{i_1}|\boldsymbol{G}) &= \mathrm{Var}(\Delta_{i_1} + \eta_{i_1}|\boldsymbol{G})\\ &= \mathrm{Var}(\Delta_{i_1}|\boldsymbol{G}) + \mathrm{Var}\Big(\frac{1}{n_1}\sum_{j:X_j=1}\epsilon_{i_1,j} - \frac{1}{n_2}\sum_{j:X_j=0}\epsilon_{i_1,j}|\boldsymbol{G}\Big)\\ &= \mathrm{Var}(\Delta_{i_1}|\boldsymbol{G}) + \frac{1}{n_1} + \frac{1}{n_2}.\end{aligned} \tag{19}$$

Equation (**10**) immediately follows from equations (**18**) and (**19**).

REFERENCES

1. Goeman, J. J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics,* **23**(8), 980–987.
2. 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.
3. 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.
4. Tsai, C.-A. and Chen, J. J. (2009) Multivariate analysis of variance test for gene set analysis. *Bioinformatics,* **25**(7), 897–903.
5. 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.
6. Huang, Y.-T. and Lin, X. (2013) Gene set analysis using variance component tests. *BMC Bioinformatics,* **14**(1), 210.
7. 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. *Proc. Natl. Acad. Sci. U.S.A,* **102**(38), 13544–13549.
8. Wu, D. and Smyth, G. K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.,* **40**(17), e133–e133.
9. 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 Res.,* p. gkt660.
10. Khatri, P., Sirota, M., and Butte, A. J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS. Comput. Biol.,* **8**(2), e1002375.
11. 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. *Proc. Natl. Acad. Sci. U.S.A,* **102**(43), 15545–15550.
12. Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.,* **3, Article3**.
13. Kim, S.-Y. and Volsky, D. J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics,* **6**(1), 144.
14. Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.,* **102**(477).
15. Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Ann. Appl. Statist.,* pp. 107–129.
16. 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.
17. 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 Res.,* **37**(1), 1–13.
18. Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013) Empirical pathway analysis, without permutation. *Biostatistics,* p. kxt004.
19. Barry, W. T., Nobel, A. B., and Wright, F. A. (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Statist.,* pp. 286–315.
20. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.,* **28**(1), 27–30.
21. 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. *Nat. Genet.,* **25**(1), 25–29.
22. 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.
23. 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.
24. Smyth, G. K. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* pp. 397–420 Springer.
25. 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 Biol.,* **5**(10), R80.

26. 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.
27. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy. Stat. Soc. B Met.,* pp. 289–300.
28. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.,* pp. 1165–1188.
29. Alexa, A. and Rahnenfuhrer, J. (2010) topGO: enrichment analysis for gene ontology. *R Package Version,* **2**(0).
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. *J. Neurosci.,* **30**(16), 5702–5712.
31. 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.
32. 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. *Hum. Mol. Genet.,* **19**(22), 4373–4384.
33. 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. *Hum. Mol. Genet.,* p. ddq322.
34. 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.