Estimating test statistics correlation from sample correlation

May 13, 2016

1 Introduction

In gene expression experiments, inter-gene correlations are commonly observed in expression data [4, 14, 1, 5, 15, 10, 17, 8]. The key task of expression analysis is to detect differentially expressed (DE) genes. One common feature of such DE detection is that a summary statistic is calculated for each gene to measure the magnitude of DE. The test statistics are often of familiar form, for example, they may come from two-sample comparison or experimental design based regression models. However, those test statistics are likely to be correlated, since their corresponding expression levels are correlated. This paper concerns the relation between test statistics correlations and the corresponding expression level correlations.

Why would people care about correlation between genes? The stochastic dependence of test statistics has brought methodological issues, in terms of accessing both individual genes and gene sets. The interest in examining individual genes is to find DE genes among tens of thousands of candidates. Multiple hypothesis testing procedures, such as *false discovery rate* (FDR) [2] and *q-value* [15], are therefore needed. In many cases, such techniques work only when test statistics are independent [2] or have positive regression dependency [3]. The goal of evaluating gene sets is to find molecular pathways or gene networks that are related to the experimental condition or factors of interest. Testing a gene set is usually done by pooling the test statistics of its member genes, and may or may not involve genes not in the test set [9]. In all situations, the correlation between test statistics is a nuisance aspect, which, if not addressed appropriately, will undermine the applicability of the corresponding approaches (REF). For example, Efron [5] showed in a simulation study that for a nominal FDR of 0.1, the actual

FDR can easily vary by a factor of 10 when correlation between test statistics exists.

What are existing ways of dealing with inter-gene correlations? A number of attempts have been made to deal with issues of inter-gene correlation when testing either individual genes or gene sets. One approach is to derive certain summary statistic from correlation among test statistics and then use it in the hypothesis testing procedure. (Do I need more examples here) For testing individual genes, Efron [5] calculates the *false discovery proportion* (FDP) conditioning on some dispersion variate which is estimated from correlation among transformed test statistics. For testing gene sets, Wu and Smyth [17] estimate a *variance inflation factor* (VIF) associated with inter-gene correlation and incorporate it into their parametric/rank-based testing procedures. The same VIF is also used by Yaari et al. [18] to account for correlation in their distribution-based gene set testing procedure. Another approach is to permute the labels of biological samples. Sample permutation generates the null distribution of test statistic for each gene. This type of permutation preserves underlying correlation structure between genes, and thus protect the test against such correlations (REF, FDR related and gene set test related). However, sample permutation method has an extra assumption, which states that the test statistics always follow the distribution they have under complete null that no gene is DE [6]. In other words, this assumption expects that the distribution of test statistics under the null is not affected by the presence of non-null cases. The *gene set enrichment analysis* (GSEA) procedure [16] falls into this category.

Key question: Are expression level correlations the same as test statistics correlation?
The first approach requires that the correlations between test statistics are known or at least can be estimated from the data. Without replicating the experiment, however, there's no way to obtain the correlation structure of test statistics because only a single test statistic is available for each gene. In the case of one-sided test (e.g., two sample t-test), one possible choice is to use sample correlations (after gene treatment effects nullified) to represent correlations among test statistics, as is done by Barry et al. [1], Efron [5], Wu and Smyth [17]. In all of the three works, it is shown by simulation only the equivalence (in terms of either distribution or numerical summarization) of sample correlation coefficient and test statistics correlation coefficient. Efron [5] estimates the distribution of z-value (transformed from corresponding two sample t-test statistics) correlation by sample correlation. Barry et al. [1] show by Monte Carlo simulation of gene expression data that a nearly linear relationship holds between test statistic correlation and sample correlation for several types of test statistic. It has, to the best of our knowledge, not

yet been fully explored in the context of two group comparison.

What did we find

In this work, we investigated the effect of testing procedures on inter-gene correlation structure regarding two group comparison. Theoretically, we proved that for two sample z-test, there is a perfect positive correlation between sample correlation coefficient r_{sample} and test statistics correlation $r_{\text{statistic}}$. For two sample t-test, the equivalence does not hold in general for $r_{\text{statistic}}$ and r_{sample}, unless all the test are true null (no DE). We demonstrated by simulation that under the null, such equivalence also holds for two group comparison of Poisson regression.

Relevant but different work

A relevant research was done by Qiu et al. [14], in which they studied the effect of different normalization procedures on the inter-gene correlation structure for microarray data. They randomly assigned 330 arrays into 15 pairs, each containing 22 arrays within each array 12558 genes. Then 15 t-statistics were calculated for each gene to mimic 15 two-sample comparisons under null hypothesis of no DE. They compared the histogram of t-statistics correlation for different normalization algorithms, and concluded that the normalization procedures are unable to completely remove the correlation between the test statistics.

2 General setup

2.1 define what do we mean by correlation

Correlation is a statistical quantity used to assess a possible linear relationship between two random variables or two sets of data sets. The degree of correlation is measured by *correlation coefficient*, a scaler taking values on the interval $[-1, 1]$. Correlation coefficient of $+1$ (-1) indicates perfect positive (negative dependence), while correlation coefficient of 0 implies no linear relationship between two random variables. Larger correlation coefficient (in absolute value) corresponds to stronger linear correlation. There are many ways to look at the correlation coefficient , many of which are special cases of Pearson's correlation coefficient [12]. For example, the *Kendall tau rank correlation coefficient* is computed as Pearson's correlation coefficient between the ranked variables.

Let (X, Y) be a random vector, and (x_j, y_j) its jth observation. The most familiar measure of dependence between two quantities is the *Pearson's correlation coefficient*. Following the notation of Lee Rodgers and Nicewander [12], We will restrict our interest to two types of Pearson's correlation coeffi-

cient: 1) standardized covariance, which we refer to as *population correlation*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}, \tag{1}$$

where μ_X and μ_Y are the expected values and $\sigma_X < \infty$ and $\sigma_Y < \infty$ are the population standard errors, and 2) a function of raw scores and means, which we refer to as *sample correlation*

$$r = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_i (y_j - \bar{y})^2}}, \tag{2}$$

where $(\bar{x}, \bar{y})$ is the vector of arithmetic mean of the observations. Throughout this paper, we will discuss the correlation between X and Y under bivariate settings.

We assume that the population mean of (X_j, Y_j) may differ across samples, but that the population covariance structure remains the same, that is,

$$E\begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \mu_{X,j} \\ \mu_{Y,j} \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\mu}_j, \quad \text{for } j = 1, \ldots, n \tag{3}$$

and

$$\mathrm{Cov}\begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\Sigma} \tag{4}$$

where ρ is the population correlation defined by equation (1).

In addition, we assume independence across samples (Note that independence implies 0 correlation, but not vise versa),

$$\mathrm{Cov}(X_{j_1}, X_{j_2}) = \mathrm{Cov}(Y_{j_1}, Y_{j_2}) = 0 \quad \text{for } j_1 \neq j_2 \tag{5}$$

In the context of gene expression study, the goal is to detect differential expression (DE)—whether the expression level of a gene is significantly correlated with treatment or experimental variables. Let $\boldsymbol{a} := (a_1, \ldots, a_n)^T$ be a vector for a contrast of interest, then DE detection for gene X can be statistically formulated as

$$H_0 : \boldsymbol{a}^T\boldsymbol{\mu}_X = d_X \ \textit{ Verses } H_1 : \boldsymbol{a}^T\boldsymbol{\mu}_X \neq d_X, \tag{6}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{\mu}_X = (\mu_{X,1}, \ldots, \mu_{X,n})^T$. This hypothesis testing procedure usually results in a "t-test similar" test statistic, in which the numerator is a linear combination of $\boldsymbol{X}$ and the denominator is its standard error. Without a loss of generality, we express the test statistics as follows

$$T_X = \frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}, \quad T_Y = \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y} \tag{7}$$

S_X and S_Y are the standard error for $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{a}^T\boldsymbol{Y}$ respectively.

Our main goal is to explore the relationship between population correlation (equation (1)) for the test statistics

$$\rho_T(n) = \text{Corr}(T_X, T_Y), \tag{8}$$

and that for their corresponding expression level

$$\rho = \text{Corr}(X, Y). \tag{9}$$

We will examine ???HOW MANY??? different test statistics having the form of equation (7).

3 Results

In this section we present the exact expression of statistics correlation coefficient for two sample t-test. In the first part, we conclude theoretically that test statistics correlation and sample correlation are perfect positive dependent for two sample z-test, but that is not always true for two sample t-test. In the second part, we simulate four different cases where test statistics correlation $r_{\text{statistics}}$ may be very different from true correlation ρ or sample correlation r_{sample}.

3.1 Theory

Theorem 1 *Let $(X_j, Y_j), j = 1, \ldots, n$ be independent random vectors with mean and covariance structures specified in equations (3) and (4). If $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ is independent of (S_X, S_Y), then the correlation of T_X and T_Y in equation (7) can be expressed as*

$$\rho_T(n) = \frac{\rho E(S_X^{-1}S_Y^{-1}) + \frac{\boldsymbol{a}^T\boldsymbol{\mu}_X \cdot \boldsymbol{a}^T\boldsymbol{\mu}_Y}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}} Cov(S_X^{-1}, S_Y^{-1})}{\sqrt{\left[E(S_X^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_X^{-1})\right]\left[E(S_Y^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_Y^{-1})\right]}} \tag{10}$$

Proof: Since samples are independent, we have

$$\begin{aligned}
\text{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) &= \boldsymbol{a}^T\text{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{a} = \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}, \\
\text{Var}(\boldsymbol{a}^T\boldsymbol{X}) &= \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}, \\
E(\boldsymbol{a}^T\boldsymbol{X})^2 &= (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2 + \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}, \\
E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})] &= E(\boldsymbol{a}^T\boldsymbol{X})E(\boldsymbol{a}^T\boldsymbol{Y}) + \text{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) \\
&= (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}
\end{aligned} \tag{11}$$

Note that since S_X is independent of S_X, we have

$$\begin{aligned}\mathrm{Var}(T_X) &= E\left[\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)^2\right] - \left[E\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)\right]^2 \\ &= E[\boldsymbol{a}^T\boldsymbol{X}]^2 E[S_X^{-2}] - \left[E(\boldsymbol{a}^T\boldsymbol{X})\right]^2\left[E(S_X^{-1})\right]^2 \\ &= \sigma_X^2 \boldsymbol{a}^T\boldsymbol{a} E(S_X^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2 \mathrm{Var}(S_X^{-1})\end{aligned} \tag{12}$$

Similarly,

$$\mathrm{Var}(T_Y) = \sigma_Y^2 \boldsymbol{a}^T\boldsymbol{a} E(S_Y^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2 \mathrm{Var}(S_Y^{-1}) \tag{13}$$

and

$$\begin{aligned}&\mathrm{Cov}(T_X, T_Y) = E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}} \cdot \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right] - E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}}\right] E\left[\frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right] \\ &= E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})] \cdot E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) E[S_X^{-1}] E[S_Y^{-1}] \\ &= [(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y \boldsymbol{a}^T\boldsymbol{a}] E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) E[S_X^{-1}] E[S_Y^{-1}]\end{aligned} \tag{14}$$

The result follows by plugging equations (11)-(14) into equation (1).

corollary 1 *For any non zero* $\boldsymbol{a}$, $\rho_T = \rho$ *if* S_X *and* S_Y *are constant with respect to* $\boldsymbol{X}, \boldsymbol{Y}$.

Proof: When S_X and S_Y are constants, $\mathrm{Cov}(S_X^{-1}, S_Y^{-1})$, $\mathrm{Var}(S_X^{-1})$ and $\mathrm{Var}(S_Y^{-1})$ are all 0, and equation (10) reduces to

$$\rho_T(n) = \frac{\rho E(S_X^{-1}S_Y^{-1})}{\sqrt{E(S_X^{-2})E(S_Y^{-2})}} = \rho. \tag{15}$$

Note: Corollary 1 states that test statistics correlation and expression level correlation are equal under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$. In a two group comparison, Let $n = n_1 + n_2$ where n_1, n_2 are the sample sizes for the two groups, and

$$\boldsymbol{a} = (\underbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}_{n_1}, \underbrace{-\frac{1}{n_2}, \ldots, -\frac{1}{n_2}}_{n_2})^T \tag{16}$$

be a contrast. If we set $S_X = 1$, then T_X corresponds to mean difference between the treatment and the control group; instead, if $S_X = \sigma_X\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where σ_X is known, then T_X corresponds to the statistic for two sample z-test. Therefore, $\rho_T = \rho$ if we use mean difference or z-value as test statistic.

Next, if we assume that (S_X, S_Y) is a function of $(\boldsymbol{X}, \boldsymbol{Y})$, then the test statistics correlation in equation (10) can be expressed as

$$\rho_T(n) = \frac{\frac{E(S_X^{-1}S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}\rho + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}}\rho_s}{\sqrt{\left[\frac{E(S_X^{-2})}{\mathrm{Var}(S_X^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}}\right]\left[\frac{E(S_Y^{-2})}{\mathrm{Var}(S_Y^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}}\right]}} \tag{17}$$

where

$$\rho_s = \frac{\mathrm{Cov}(S_X^{-1}, S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}} \tag{18}$$

SAY SOMETHING HERE....

In gene expression analysis, a commonly used statistic is the two sample t-statistic (REF). Let $(\mu_{X,j}, \mu_{Y,j}) = (\mu_X, \mu_Y)$ for $j = 1, \ldots, n_1$ and $(\mu_{X,j}, \mu_{Y,j}) = (\mu_X + \Delta_X, \mu_Y + \Delta_Y)$ for $j = n_1 + 1, \ldots, n_1 + n_2$. In the case of two sample t-test with equal variance, $\boldsymbol{a}^T \boldsymbol{X} = \bar{X}_1 - \bar{X}_2$ and $S_X = S_{P,X}\sqrt{1/n_1 + 1/n_2}$ where $S_{P,X}$ is the pooled variance. Denote

$$\begin{aligned} &\boldsymbol{a}^T \boldsymbol{X} = \bar{X}_1 - \bar{X}_2 \stackrel{def}{=} U_X, \quad \boldsymbol{a}^T \boldsymbol{Y} = \bar{Y}_1 - \bar{Y}_2 \stackrel{def}{=} U_Y, \\ &S_{P,i}^2 = \frac{(n_1 - 1)S_{i,1}^2 + (n_2 - 1)S_{i,2}^2}{n_1 + n_2 - 2}, \quad i = X, Y. \end{aligned} \tag{19}$$

where $\boldsymbol{a}$ is defined in equation (16) and $S_{i,1}^2$ and $S_{i,2}^2$ are sample variances for the two groups for gene i. Under two sample t-test with equal variance, we have the following theorem.

Theorem 2 *Let $(X_i, Y_i), i = 1, \ldots, n_1 + n_2$ follow a bivariate normal distribution with mean specified in equation (3) and covariance in (4). If T_X and T_Y are statistics for equal-variance two-sample t-test, then*

$$Corr(T_X, T_Y) = \frac{\frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} C\rho_s + \rho B + \rho_s \rho (A - B)}{\sqrt{\left[\frac{\Delta_X^2}{\sigma_X^2} C + A\right]\left[\frac{\Delta_Y^2}{\sigma_X^2} C + A\right]}} \tag{20}$$

where

$$\begin{aligned} &A = \frac{n_1 + n_2 - 2}{n_1 + n_2 - 4}, \quad B = \frac{(\frac{n_1+n_2-2}{2})\Gamma^2(\frac{n_1+n_2-4}{2} + \frac{1}{2})}{\Gamma^2(\frac{n_1+n_2-2}{2})}, \\ &\rho_s = Corr(\frac{1}{S_X}, \frac{1}{S_Y}), \quad C = \frac{(n_1 + n_2)(A - B)}{(2 + n_1 n_2^{-1} + n_1 n_2^{-1})}. \end{aligned} \tag{21}$$

The proof of Theorem 2 is presented in Section 4. Next we give the limit of $\mathrm{Corr}(T_X, T_Y)$.

Theorem 3 *If there exists positive constant M_1 and M_2, such that $M_1 \leq n_1 n_2^{-1} \leq M_2$, then*

$$\rho_T = \lim_{n_1+n_2 \to \infty} Corr(T_X, T_Y) = \frac{\rho + \beta \frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} \rho_s}{\sqrt{\left[1 + \beta \frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta \frac{\Delta_Y^2}{\sigma_Y^2}\right]}} \tag{22}$$

where ρ_s is defined in equation (21) and $\beta = \lim_{n_1+n_2 \to \infty} C = (4 + 2n_1^{-1}n_2 + 2n_1 n_2^{-1})^{-1}$.

Theorem 3 says that as long as n_1 and n_2 grow proportionally, the limit of $\text{Corr}(T_X, T_Y)$ is a function of population covariance $\mathbf{\Sigma}$, the true mean difference $\mathbf{\Delta}$ and the ratio n_1/n_2. We have the following observations:

1. If both test are true null (i.e., $\mathbf{\Delta} = \mathbf{0}$), then $\rho_T = \rho$.
2. If one test is true null, then ρ_T is proportional to and smaller in absolute value than ρ. For example, when $\Delta_X = 0$, $\rho_T = \rho/\sqrt{1 + \beta\frac{\Delta_Y^2}{\sigma_Y^2}}$.
3. If both tests are true alternative (i.e., $\mathbf{\Delta} \neq \mathbf{0}$), then $\rho_T \neq \rho$ in general.

However, it should be noted that if $n_1/n_2 \to 0$ or ∞, then $\beta = 0$ and we have $\rho_T = \rho$. That is, when sample size of one group is dominant, $\text{Corr}(T_X, T_Y)$ will converge to ρ regardless of whether the tests are under the null or not.

Depending on the underlying value of $\mathbf{\Delta}$ (DE or not DE, up-regulated or down-regulated if DE) and covariance $\mathbf{\Sigma}$, ρ_T might be far from ρ in different ways. Next we will discuss the relationship between ρ_T and ρ under several scenarios.

We show via simulation (see Figure 1) that for ρ growing from -1 to 1, ρ_s in equation (21) has a "U" shape whose minimum is located near $\rho = 0$, and

$$0 \leq r_s \leq |\rho| \quad \text{ONLY BASED ON SIMULATION} \tag{23}$$

which is useful in comparing ρ_T and ρ. For $\rho < 0$

1. if $\Delta_X\Delta_Y > 0$, then gene X and gene Y are DE in the same direction (both up-regulated or both down-regulated), then

$$\rho_T = \frac{\rho + \beta\frac{\Delta_X\Delta_Y}{\sigma_X\sigma_Y}\rho_s}{\sqrt{\left[1 + \beta\frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta\frac{\Delta_Y^2}{\sigma_Y^2}\right]}} > \frac{\rho}{\sqrt{\left[1 + \beta\frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta\frac{\Delta_Y^2}{\sigma_Y^2}\right]}} > \rho \tag{24}$$

2. if $\Delta_X\Delta_Y < 0$, then gene X and gene Y are DE in different directions (one up-regulated and the other down-regulated), then by $\rho_s < -\rho$,

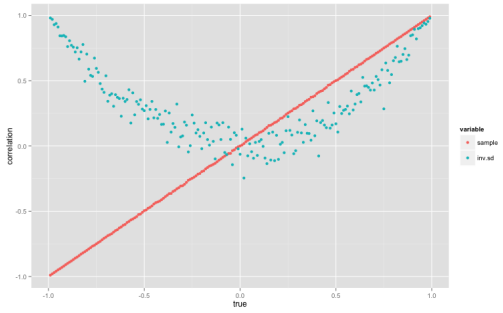
$$\rho(T_X, T_Y) = \frac{\rho + \beta\frac{\Delta_X\Delta_Y}{\sigma_X\sigma_Y}\rho_\text{s}}{\sqrt{\left[1 + \beta\frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta\frac{\Delta_Y^2}{\sigma_Y^2}\right]}} > \rho\frac{1 - \beta\frac{\Delta_1\Delta_2}{\sigma_1\sigma_2}}{\sqrt{\left[1 + \beta\frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta\frac{\Delta_Y^2}{\sigma_Y^2}\right]}} > \rho$$

3. if $\Delta_X\Delta_Y = 0$, then one is DE but the other is not. Suppose gene X is not DE, then

$$\rho(T_X, T_Y) = \frac{\rho}{\sqrt{\left[1 + \beta\frac{\Delta_X^2}{\sigma_X^2}\right]}} > \rho$$

Therefore in any case, $\rho_T \geq \rho$ when $\rho < 0$. Similarly it can be shown that for $\rho > 0$, $\rho_T \leq \rho$. In simple words, T_X and T_Y are "less" correlated than their corresponding expression values are.

Figure 1: $\text{Corr}(S_X^{-1}, S_Y^{-1})$ against ρ



3.2 Simulation

We perform simulations to evaluate the correlations between test statistics and those between expression levels under two sample t-test. We simulate the expression data from normal distributions. Specifically, we let (X, Y) be the expression levels of genes X and Y, and

$$\begin{aligned} \begin{pmatrix} X_{j_1} \\ Y_{j_1} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sqrt{1 \cdot 3} \\ \rho\sqrt{1 \cdot 3} & 3 \end{pmatrix} \right] \\ \begin{pmatrix} X_{j_2} \\ Y_{j_2} \end{pmatrix} &\sim N\left[\begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho\sqrt{1 \cdot 3} \\ \rho\sqrt{1 \cdot 3} & 3 \end{pmatrix} \right] \end{aligned} \tag{25}$$

where $j_1 = 1, \ldots, n_1$ and $j_2 = n_1 + 1, \ldots, n_1 + n_2$. In this simulation setting, we set both n_1 and n_2 to be 100. For each given ρ, we consider these $n = 200$ pairs of (X, Y) as observations from one *simulated* experiment. Out of this experiment, we calculate $q = (T_X, T_Y, r_{XY})$ where T_X and T_Y are the test statistics for gene X and gene Y respectively using two-sample t-test for equal variance procedure, and r_{XY} is the sample correlation after the treatment effect is removed. We replicate the simulated experiment for $B = 1000$ times, resulting in a matrix $\boldsymbol{Q}_{1000\times 3}$. We take the correlation between the first and the second columns of $\boldsymbol{Q}$ as an estimate for test statistics correlation $r_{\text{statistics}}$, and the mean of the third column as an estimate of sample correlation r_{sample}. Fisher [7] proved that sample correlation is a consistent estimator for underlying true correlation, therefore $r_{\text{statistics}}$ and r_{sample} should reflect the true correlation between T_X and T_Y and that between X and Y

respectively. We increase ρ from -0.99 to 0.99 by fixed step size 0.01, and examine the relationship between $r_{\text{statistics}}$ and r_{sample} under four different cases:

a) No DE genes (i.e., $\Delta_X = \Delta_Y = 0$);

b) One gene is DE and the other is not (i.e, only one of Δ_X and Δ_Y is 0); in the simulation we set $\Delta_X = 0$ and $\Delta_Y = 5$;

c) DE towards the same direction (i.e., $\Delta_X \Delta_Y > 0$); in the simulation we set $\Delta_X = 2$ and $\Delta_Y = 5$;

d) DE towards opposite directions (i.e., $\Delta_X \Delta_Y < 0$); in the simulation we set $\Delta_X = 2$ and $\Delta_Y = -5$.

In Figure 2, we plot $r_{\text{statistics}}$ and r_{sample} against the underlying true population correlation ρ. Note that in all cases, while r_{sample} is a consistent estimator of ρ, $r_{\text{statistics}}$ might be very different from ρ and thus from r_{sample}. In case a) where no gene is DE, $r_{\text{statistics}}$ and r_{sample} are almost equal, and both converge the true correlation ρ. However, as long as DE effect exists, there is a discrepancy between $r_{\text{statistics}}$ and ρ. In case b) where only one gene is DE, the magnitude of $r_{\text{statistics}}$ is proportional to, and smaller in absolute value than ρ. It is more interesting to note that $r_{\text{statistics}}$ is not monotone with respect to ρ when both genes are DE. If genes are DE towards the same direction as in the case of c), $r_{\text{statistics}}$ first decreases from a positive value to 0, and continues to decrease until it reaches the minimum (a negative value), and then gradually increases to 1, as ρ grows from -1 to 1. When genes are DE towards opposite directions like in case d), however, the trend is reversed from that in the case of c). This set of simulation results is reflected in the test statistics correlation formula of equation (22).

4 Method

Lemma 1 *Sample correlation coefficient is a consistent estimator for ρ,*

$$\sqrt{n}(r_{sample} - \rho) \xrightarrow{D} N\left(0, (1-\rho^2)^2\right).$$

The proof of lemma 1 can be found in Fisher [7].

To prove Theorem 2, it is useful to note that $\boldsymbol{U}$ is independent of $\boldsymbol{S}$, following from Lemmas 2 and 3.

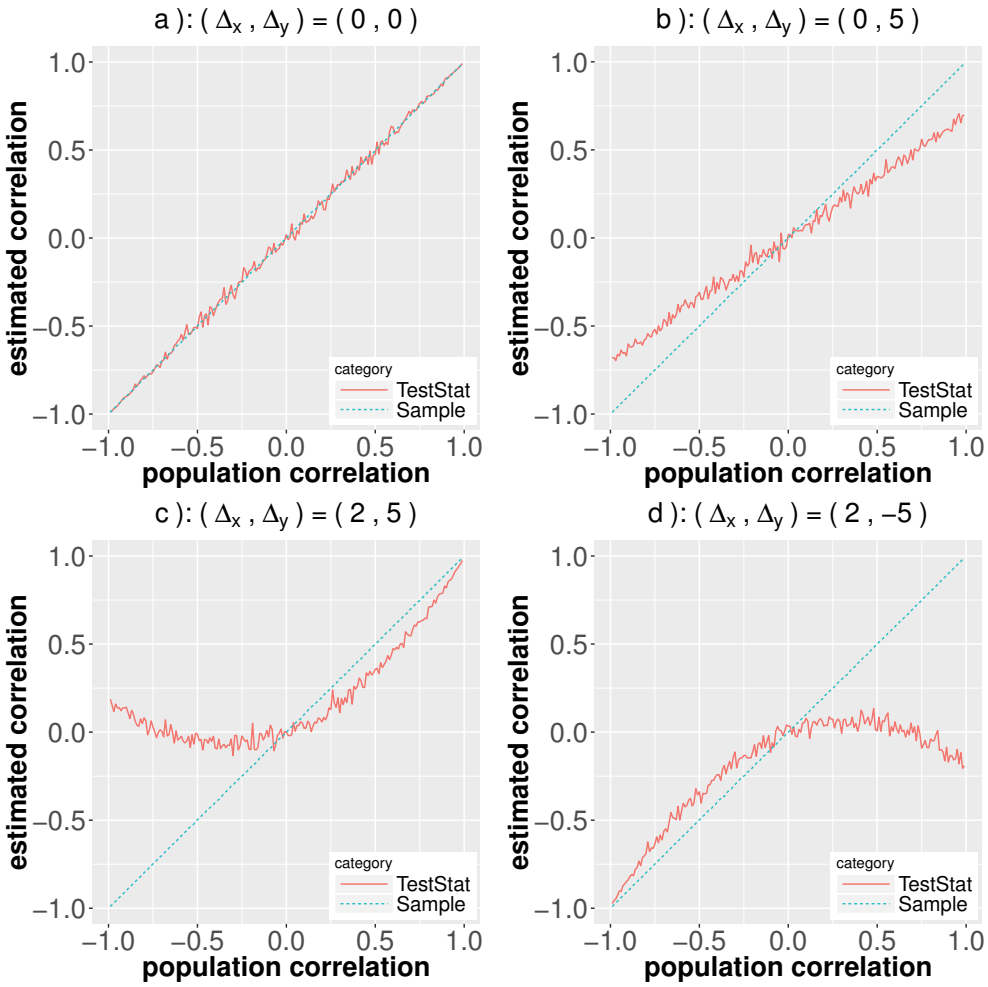


Figure 2: Plots for estimates of sample/test statistics correlation against true population correlations. For each of the simulation settings a)–d), the test statistics are calculated using two sample t-test with equal variance, and the correlations are calculated by equation (2).

Lemma 2 *Let* $(X_j, Y_j), j = 1 \ldots, n$ *be independent random variables satisfying equation (5), then* $\boldsymbol{W} = (W_X, W_Y) = (\frac{(n-1)S_X^2}{\sigma_X^2}, \frac{(n-1)S_Y^2}{\sigma_Y^2})$ *follows a* ***bivariate chi square distribution*** *with density*

$$f(w_x, w_y) = \frac{2^{-n}(w_x w_y)^{(n-3)/2} e^{-\frac{w_x + w_y}{2(1-\rho^2)}}}{\sqrt{\pi}\Gamma(\frac{n}{2})(1-\rho^2)^{(n-1)/2}} \times \sum_{k=0}^{\infty}[1 + (-1)^k] \left(\frac{\rho\sqrt{w_x w_y}}{1-\rho^2}\right)^k \frac{\Gamma(\frac{k+1}{2})}{k!\Gamma(\frac{k+n}{2})} \tag{26}$$

for $n > 3$ *and* $-1 < \rho < 1$.

For proof of Lemma 2, interested readers are referred to Joarder [11]. It immediately follows from Lemma 2 that $\boldsymbol{W}_1 = (\frac{(n_1-1)S^2_{X,1}}{\sigma^2_X}, \frac{(n_1-1)S^2_{Y,1}}{\sigma^2_Y})$ follows bivariate chi-square distribution with degree of freedom $n_1 - 1$, where $S_{X,1}$ and $S_{X,2}$ are defined in equation (**??**). Similarly, $\boldsymbol{W}_2 = (\frac{(n_2-1)S^2_{X,2}}{\sigma^2_X}, \frac{(n_2-1)S^2_{Y,2}}{\sigma^2_Y})$ follows a bivariate chi-square distribution with degree of freedom n_2-1. Note that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independent since the samples are independent.

Lemma 3 $\boldsymbol{U} = (U_X, U_Y)$ *is independent of* $\boldsymbol{S} = (S_X, S_Y)$*, where* $\boldsymbol{U}$ *and* $\boldsymbol{S}$ *are defined in equation (19).*

Proof: By Lemma 2, the density function of $\boldsymbol{W}_1+\boldsymbol{W}_2$ only involves $\sigma^2_X, \sigma^2_Y, \rho$ and sample size n_1, n_2, therefore we can denote its density by some function $g(\sigma^2_X, \sigma^2_Y, \rho, n_1 + n_2)$. Note that $\boldsymbol{S}^2 = \frac{(\sigma^2_X,\ \sigma^2_Y)}{n_1+n_2-2}(\boldsymbol{W}_1 + \boldsymbol{W}_2)^T$ is a linear transformation of $\boldsymbol{W}_1 + \boldsymbol{W}_2$, so its density also can be expressed in terms of $\sigma^2_1, \sigma^2_2, \rho, n_1, n_2$. Therefore $\boldsymbol{S} = (S_X, S_Y)$ is an ancillary statistic for $\boldsymbol{\Delta}$. On the other hand, it can be shown that $\boldsymbol{U} = (U_X, U_Y)$ is a complete sufficient statistic for $\boldsymbol{\Delta}$. It follows by Basu's theorem that $\boldsymbol{U}$ and $\boldsymbol{S}$ are independent.

Lemma 3 implies that $U_X U_Y$ is also independent of $\frac{1}{S_X S_Y}$, and therefore $E(\frac{U_X}{S_X} \cdot \frac{U_Y}{S_Y})$ can be expressed as $E(U_X U_Y)E(\frac{1}{S_X S_Y})$. Additionally, if we know $\mathrm{Corr}(\frac{1}{S_X}, \frac{1}{S_Y})$, then the t-test statistics correlation can be accurately represented.

Proof of theorem (2)
First note that

$$\begin{aligned}\mathrm{Cov}(T_X, T_Y) &= E(T_X T_Y) - E(T_X)E(T_Y) \\ &= \frac{1}{c_0^2}\left[E(U_X U_Y)E(\frac{1}{S_X S_Y}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})\right] \qquad \text{(by lemma 3)}\end{aligned}$$

where $c_0 = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and $\mathrm{Var}(T_X) = \mathrm{Var}(\frac{U_X}{c_0 S_X}) = \frac{1}{c_0^2}\mathrm{Var}(\frac{U_X}{S_X})$. Note that

$$\begin{aligned}\mathrm{Corr}(T_X, T_Y) &= \frac{\mathrm{Cov}(T_X, T_Y)}{\sqrt{\mathrm{Var}(T_X)\mathrm{Var}(T_Y)}} \\ &= \frac{E(U_X U_Y)E(\frac{1}{S_X S_Y}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})}{\sqrt{\mathrm{Var}(\frac{U_X}{S_X})\mathrm{Var}(\frac{U_Y}{S_Y})}}\end{aligned} \tag{27}$$

We need to calculate $E(U_X U_Y)$, $E(\frac{1}{S_X S_Y})$, $E(\frac{U_i}{S_i})$ and $\mathrm{Var}(\frac{U_i}{S_i})$ for $i = X, Y$.

1. Note that $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1}+\frac{1}{n_2})\right), i = X, Y.$

$$\begin{aligned} E(U_X U_X) &= \mathrm{Cov}(U_X, U_Y) + E(U_X)E(U_Y) \\ &= \rho\sigma_X\sigma_Y\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \Delta_X\Delta_Y \end{aligned} \tag{28}$$

2. Since $\frac{(n_1-1)S_X^2}{\sigma_X^2}$ and $\frac{(n_2-1)S_Y^2}{\sigma_Y^2}$ are independent and follow $\chi^2(n_1-1)$ and $\chi^2(n_2-1)$ respectively, , we have $W_X = \frac{(n_1+n_2-2)S_X^2}{\sigma_X^2} \sim \chi^2(n_1+n_2-2)$. It can be shown that

$$E(W_X^k) = \frac{2^k\Gamma(\frac{n_1+n_2-2}{2}+k)}{\Gamma(\frac{n_1+n_2-2}{2})}$$

Therefore

$$E\left(\frac{1}{S_X}\right) = \frac{\sqrt{B}}{\sigma_X}, \quad \mathrm{Var}\left(\frac{1}{S_X}\right) = \frac{A-B}{\sigma_X^2} \tag{29}$$

Note that $\rho_s = \mathrm{Corr}(\frac{1}{S_X}, \frac{1}{S_Y})$, we have

$$\begin{aligned} E(\frac{1}{S_X S_Y}) &= E(\frac{1}{S_X})E(\frac{1}{S_Y}) + \rho_s\sqrt{\mathrm{Var}(\frac{1}{S_X})\mathrm{Var}(\frac{1}{S_Y})} \\ &= \frac{B}{\sigma_X\sigma_Y} + \rho_s\frac{A-B}{\sigma_X\sigma_Y} \end{aligned} \tag{30}$$

3. $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1}+\frac{1}{n_2})\right)$ and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2} \sim \chi^2(n_1+n_2-2)$ and by Lemma 3 U_i and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}$ are independent for $i = X, Y$, we have

$$\frac{\frac{U_i-\Delta_i}{\sigma_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}/(n_1+n_2-2)} = \frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{31}$$

It follows from

$$E\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = 0, \quad \mathrm{Var}\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{n_1+n_2-2}{n_1+n_2-4} \tag{32}$$

that

$$\begin{aligned} E\left(\frac{U_i}{S_i}\right) &= \frac{\Delta_i}{\sigma_i}\sqrt{B} \\ \mathrm{Var}\left(\frac{U_i}{S_i}\right) &= A\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{\Delta_i^2}{\sigma_i^2}(A-B) \end{aligned} \tag{33}$$

Finally, the test statistics correlation (20) is obtained by plugging equations (28–33) into equation (27).

Up to now we have obtained an exact expression for $\mathrm{Corr}(T_X, T_Y)$, which depends not only on the sample size n_1 and n_2, but also on Δ/σ, the relative magnitude of DE. The rest of this section discusses asymptotic property of $\mathrm{Corr}(T_X, T_Y)$ for large sample size.

Lemma 4 *If there exists a positive number M, such that $n_1 n_2^{-1} \leq M$ and $n_1 n_2^{-1} \leq M$, then the following results hold:*

1. $\lim_{n_1+n_2\to\infty} A = 1.$
2. $\lim_{n_1+n_2\to\infty} B = 1.$
3. $\lim_{n_1+n_2\to\infty} C = \beta.$

where A, B and C are defined in equation (21), and $\beta = (4 + n_1 n_2^{-1} + n_1^{-1} n_2)^{-1}$.

Proof: Note that

$$B = \begin{cases} \frac{(k-1)\Gamma^2(k-\frac{3}{2})}{\Gamma^2(k-1)}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})\Gamma^2(k-1)}{\Gamma^2(k-\frac{1}{2})}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{34}$$

We will use second order Stirling's formula,

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k (1 + \frac{1}{12k}) \tag{35}$$

Using Stirling's formula (35) and $\Gamma(k+\frac{1}{2}) = \frac{(2k)!}{4^k k!}\sqrt{\pi}$, it can be shown that

$$B \approx \begin{cases} \frac{(k-1)(k-2)(k-2+\frac{1}{24})^2}{(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})(k-1+\frac{1}{12})^4}{(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{36}$$

It can also be shown using equation (36) that

$$A - B \approx \begin{cases} \frac{\frac{1}{4}(k-1)(k-2)^3+o((k-2)^4)}{(k-2)(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{\frac{1}{4}(k-1)^3(k-\frac{1}{2})(k-3)+o((k-1)^4)}{(k-\frac{3}{2})(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{37}$$

And the results immediately follow.

5 Conclusion

State the major findings

This article discusses the relationship between sample correlation coefficients r_{sample} (after treatment effects removed) and test statistics correlation $r_{\text{statistics}}$ in a two group comparison setting. We proved that under normal distribution assumption, $r_{\text{statistics}}$ and r_{sample} have a perfect positive correlation for two sample z test. However, for two sample t-test this correspondence does not hold in general, unless the null in (**??**) is true for all the tests considered. The results for two sample t-test can be applied to the case of two group mean comparison under Poisson regression, as shown by simulation. Consequently, that estimating $r_{\text{statistics}}$ by r_{sample} after nullifying treatment effects can not be taken for granted.

State the practical meaningness of the findings

In gene expression analysis, cares need to be taken when estimating test statistics correlation from sample correlation. For microarray data, two sample t test ([5], [1]) or its moderated version [17] are used in detecting DE, with $r_{\text{statistics}}$ estimated from sample correlation to adjust for inter-gene correlation. Our study shows, however, that for DE genes, $r_{\text{statistics}}$ may be either overestimated if two genes are positively correlated, or underestimated if two genes are negatively correlated. If we believe that most genes are positively correlated (if any) and that there are true DE genes, then the VIF factor may be overestimated in [17], which may result in conservative test for enrichment analysis; the variance of $r_{\text{statistics}}$ may also be overestimated in [5], which leads to larger variation in estimating conditional FDP. The situation may be more complicated for RNA-Seq data, which are counts in nature and therefore need to be modeled by more sophisticated regression tools (e.g. logistic regression, negative binomial regression, etc.).

Acknowledge the study's limitations

One assumption yet to be justified

In the context of two sample t-test, the simulation results agree with our theoretical conclusion, assuming that $0 \leq r_s \leq |\rho|$ in (23) is true. Our simulation does suggest

$$r_s = \rho^2, \tag{38}$$

as shown in figure (1). If (38) can be justified theoretically, it is possible to approximate the true value of $\rho(T_1, T_2)$, which will correct the bias of estimating $r_{\text{statistics}}$ by r_{sample}. Another remaining challenge is to assess the relationship of $r_{\text{statistics}}$ and ρ for non-normal distributions, or for other hypothesis testing under different regression models (e.g., negative binomial regression).

References

[1] 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.

[2] 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.

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

[4] Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*.

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

[6] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

[7] Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, pages 507–521.

[8] 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.

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

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

[11] Joarder, A. H. (2009). Moments of the product and ratio of two correlated chi-square variables. *Statistical Papers*, 50(3):581–592.

[12] Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

[13] Madsen, L. and Birkes, D. (2013). Simulating dependent discrete data. *Journal of Statistical Computation and Simulation*, 83(4):677–691.

[14] Qiu, X., Brooks, A. I., Klebanov, L., and Yakovlev, A. (2005). The effects of normalization on the correlation structure of microarray data. *BMC bioinformatics*, 6(1):120.

[15] Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035.

[16] 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.

[17] 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.

[18] 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.