Estimating test statistics correlation from sample correlation

May 18, 2016

1 Introduction

In gene expression experiments, inter-gene correlations are commonly observed in expression data [4, 13, 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. [13], 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. Throughout this paper, we will discuss the correlation between X and Y under bivariate settings.

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

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

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.

Let (X_j, Y_j), be a bivariate random variable representing two features of sample $j = 1, \ldots, m$, and (x_j, y_j) the corresponding realization. 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,

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

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 $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ and $\boldsymbol{\mu}_X = (\mu_{X,1}, \ldots, \mu_{X,m})^T$. DE detection for gene Y can be obtained by applying the same contrast to $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ (simply replacing the subscript X by Y in equation (6)). 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

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

where 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

$$\rho_T(m) = \mathrm{Corr}(T_X, T_Y), \tag{8}$$

and that for their corresponding expression level

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

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

3 Results

In this section we present the exact formula of test statistics correlation $\rho_T(m)$ by making some assumptions about T_X and T_Y, and show that the test statistics correlation ρ_T does not always equal to the population correlation ρ. For the case of two-group comparison, we prove that 1) if T_X (T_Y) is a linear transformation of $\boldsymbol{X}$ ($\boldsymbol{Y}$), then $\rho_T = \rho$, and that 2) if T_X (T_Y) is the two sample t-test statistic for $\boldsymbol{X}$ ($\boldsymbol{Y}$), then $|\rho_T| \leq |\rho|$. For 2), we show that the relationship between ρ_T and ρ depends on whether the hypotheses tests (equation 6) are true null or not. We perform simulations for the case of test statistics derived from two-sample t-test to illustrate our findings.

3.1 Theory

Theorem 1 *Let $(X_j, Y_j), j = 1, \ldots, m$ be independent random vectors with mean and covariance structures specified in equation (3). 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*

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

Proof: Since samples are independent, we have

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

Note that since S_X is independent of S_X, we have

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

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

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

Corollary 1 states that test statistics correlation and expression level correlation are equal under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$. However, if we assume that (S_X, S_Y) is a non-constant function of $(\boldsymbol{X}, \boldsymbol{Y})$, then the test statistics correlation in equation (10) can be expressed as

$$\rho_T(m) = \frac{\frac{E(S_X^{-1}S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}\rho + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}}\rho_s}{\sqrt{\left[\frac{E(S_X^{-2})}{\mathrm{Var}(S_X^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}}\right]\left[\frac{E(S_Y^{-2})}{\mathrm{Var}(S_Y^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}}\right]}} \tag{16}$$

where

$$\rho_s = \frac{\mathrm{Cov}(S_X^{-1}, S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}. \tag{17}$$

The correlation between test statistics $\rho_T(m)$ depends on the form of test statistics, and in general, may not converge to the population correlation ρ.

3.2 Application of Theorem 1 to two group comparisons

Many gene expression experiments are done to compare expression levels under two-treatment conditions. For the rest of this section, we discuss the relationship between ρ_T and ρ under such setting. Let $n = n_1 + n_2$ be the total number of samples, where n_1 of them are from group 1 and n_2 from group 2, and let

$$\boldsymbol{a} = (\underbrace{\frac{1}{n_1}, \dots, \frac{1}{n_1}}_{n_1}, \underbrace{-\frac{1}{n_2}, \dots, -\frac{1}{n_2}}_{n_2})^T \tag{18}$$

be the contrast of interest. The mean expression levels are specified as

$$\begin{aligned} \boldsymbol{\mu}_j &= (\mu_X, \mu_Y), \quad j = 1, \dots, n_1, \\ \boldsymbol{\mu}_j &= (\mu_X, \mu_Y)^T + (\Delta_X, \Delta_Y)^T, \quad j = n_1 + 1, \dots, n_1 + n_2. \end{aligned} \tag{19}$$

If we set $S_X = 1$, then T_X corresponds to mean difference between groups 1 and 2; 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, according to Corollary 1, $\rho_T = \rho$ if we use mean difference or z-value as test statistics.

The two sample t-statistic is also a commonly used statistic in differential expression analysis. In the case of two sample t-test with equal variance, with the contrast $\boldsymbol{a}$ defined in equation (18), the test statistic for X is

$$T_X = \frac{\bar{X}_1 - \bar{X}_2}{S_{p,X}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{20}$$

where $S_{p,X}$ is the pooled variance

$$S_{p,X}^2 = \frac{(n_1-1)S_{X,1}^2 + (n_2-1)S_{X,2}^2}{n_1+n_2-2}. \tag{21}$$

Similarly, we obtain T_Y by replacing the subscript "X" in equations (20) and (21). Under normal distribution assumption, we have the following theorem for two sample t-test with equal variance:

Theorem 2 *Let $(X_i, Y_i), i = 1, \ldots, n$ follow a bivariate normal distribution with mean specified by equations (19) and covariance $\mathbf{\Sigma}$ (see equation (3)). If T_X and T_Y are statistics for equal-variance two-sample t-test, then*

$$Corr(T_X, T_Y) = \frac{\frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} C\rho_s + \rho B + \rho_s \rho(A-B)}{\sqrt{\left[\frac{\Delta_X^2}{\sigma_X^2}C + A\right]\left[\frac{\Delta_Y^2}{\sigma_X^2}C + A\right]}} \tag{22}$$

where

$$\begin{aligned} A &= \frac{n_1+n_2-2}{n_1+n_2-4}, \quad B = \frac{\left(\frac{n_1+n_2-2}{2}\right)\Gamma^2\left(\frac{n_1+n_2-4}{2}+\frac{1}{2}\right)}{\Gamma^2\left(\frac{n_1+n_2-2}{2}\right)}, \\ \rho_s &= Corr(S_X^{-1}, S_Y^{-1}), \quad C = \frac{(n_1+n_2)(A-B)}{(2+n_1 n_2^{-1} + n_1 n_2^{-1})}. \end{aligned} \tag{23}$$

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

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

$$\rho_T = \lim_{n_1+n_2\to\infty} Corr(T_X, T_Y) = \frac{\rho(1+\beta\frac{\Delta_X\Delta_Y}{\sigma_X\sigma_Y}\rho)}{\sqrt{\left[1+\beta\frac{\Delta_X^2}{\sigma_X^2}\right]\left[1+\beta\frac{\Delta_Y^2}{\sigma_Y^2}\right]}} \tag{24}$$

where $\beta = \lim_{n_1+n_2\to\infty} C = (4 + 2n_1^{-1}n_2 + 2n_1n_2^{-1})^{-1}$.

Theorem 3 says that as long as n_1 and n_2 grow proportionally to infinity, the quantity ρ_T is a function of population correlation ρ, the signal-to-noise ratio $(\Delta_X/\sigma_X, \Delta_Y/\sigma_Y)$ and the sample 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 (i.e., $|\rho_T| < |\rho|$).

3. If both tests are true alternative (i.e., $\mathbf{\Delta} \neq \mathbf{0}$), then $\rho_T \neq \rho$ in general. Specifically,

   i) when $\Delta_X \Delta_Y > 0$ (i.e., both genes are DE towards the same direction), we have $\rho_T > \rho$ for $\rho < 0$ and $0 \leq \rho_T \leq \rho$ for $\rho \geq 0$.

   ii) when $\Delta_X \Delta_Y < 0$ (i.e., genes are DE towards different directions), we have $\rho < \rho_T < 0$ for $\rho < 0$ and $\rho_T < \rho$ for $\rho > 0$.

   Therefore in either case, we have $|\rho_T| \leq |\rho|$.

We note that $|\rho_T| \leq |\rho|$ when test statistics are derived from two sample t test with equal variance. In other words, T_X and T_Y are always "no more correlated" than X and Y are. It's also interesting to note that when both genes are DE, $\rho_T = 0$ at $\rho = -\frac{\sigma_X \sigma_Y}{\beta \Delta_X \Delta_Y}$ and $\frac{\sigma_X \sigma_Y}{\beta \Delta_X \Delta_Y} \in (-1, 1)$.

In addition, we note 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 not proportional to that of the other, $\mathrm{Corr}(T_X, T_Y)$ will converge to ρ regardless of whether the tests are under the null or not.

3.3 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 effects are 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 1, 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 of c): $r_{\text{statistics}}$ increases from -1 to a positive value and reaches its maximum (a positive value), and decreases to a negative value. This set of simulation results is reflected in the test statistics correlation formula of equation (24).

4 Method

Lemma 1 *The sample correlation coefficient r defined in equation (2) is a consistent estimator for the population correlation ρ,*

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

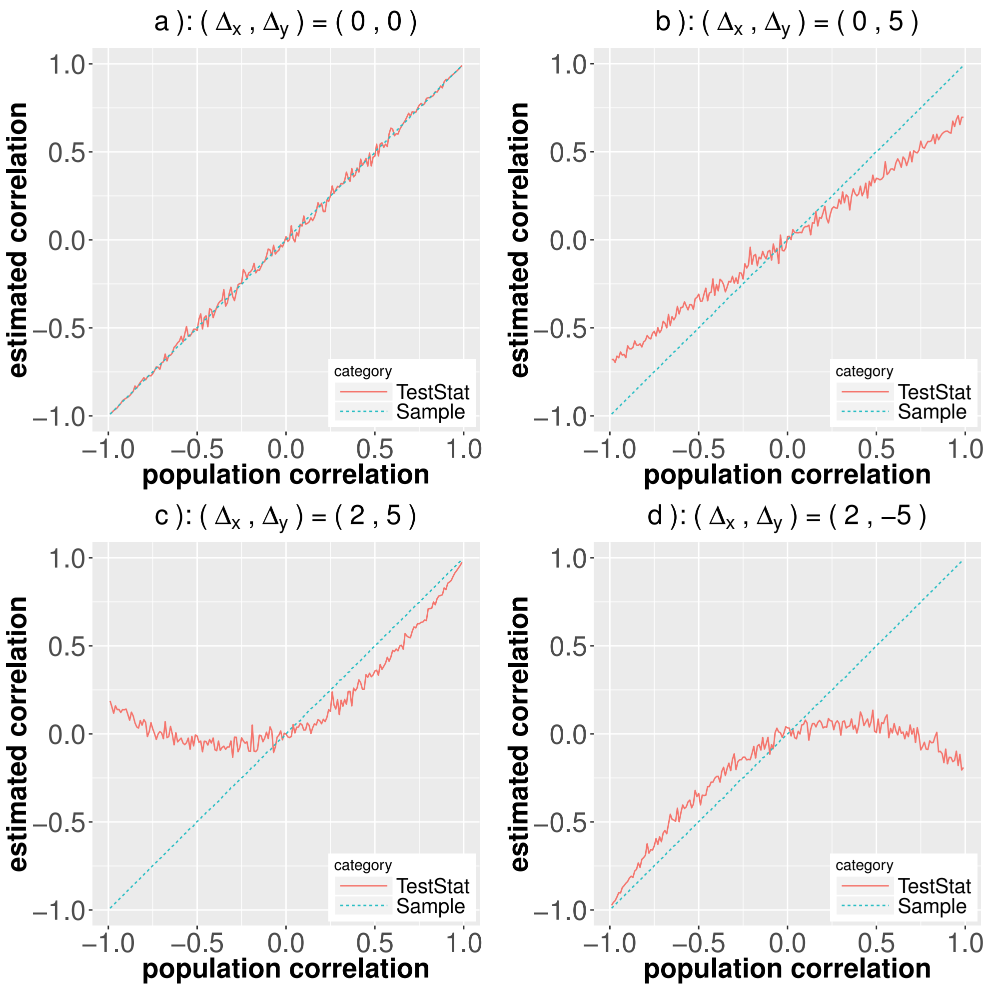

Figure 1: Plots for estimates of sample/test statistics correlation against true population correlations. For each of the simulation settings a)–d), the test statistics are calculated using two sample t-test with equal variance, and the correlations are calculated by equation (2).

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

To prove Theorem 2, it is useful to note that $\boldsymbol{U} = (\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ is independent of $\boldsymbol{S} = (S_X, S_Y)$, following from Lemmas 2 and 3.

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

bivariate chi square distribution *with density*

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

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 $n_1 - 1$. 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 $\boldsymbol{S} = (S_X, S_Y)$.*

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 $S_X^{-1}S_Y^{-1}$, and therefore $E(\frac{U_X}{S_X}\cdot\frac{U_Y}{S_Y})$ can be expressed as $E(U_X U_Y)E(S_X^{-1}S_Y^{-1})$. We can apply Theorem 1 to calculate the correlation between T_X and T_Y under two sample t-test for equal variance.

Proof of theorem 2

First note that by Lemma 3 we have

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

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

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

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

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

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

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

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

Note that $\rho_s = \text{Corr}(S_X^{-1}, S_Y^{-1})$, we have

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

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

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

that

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

Finally, the test statistics correlation (22) is obtained by plugging equations (28–33) into equation (27).

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. $\lim\limits_{n_1+n_2\to\infty} A = 1.$

2. $\lim\limits_{n_1+n_2\to\infty} B = 1.$

3. $\lim\limits_{n_1+n_2\to\infty} C = \beta.$

where A, B and C are defined in equation (23), 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.

Lemma 5 *Let $(X_j, Y_j), j = 1, \ldots, n$ be i.i.d. random variables under the two sample t-test for equal variance setting, with mean specified in equation (19) covariance structure in equation (3). Then we have $\lim_{n\to\infty} \rho_s = \rho^2$.*

Proof: Let's first look at samples $j = 1, \ldots, n_1$. Note that

$$S_{X,1}^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}(X_j - \bar{X}_1)^2 \tag{38}$$

is the *maximum likelihood estimator* (MLE) for σ_X^2. By invariance property of MLE, the pooled variance estimator

$$\begin{pmatrix} S_X^2 \\ S_Y^2 \end{pmatrix} = a_1 \begin{pmatrix} S_{X,1}^2 \\ S_{Y,1}^2 \end{pmatrix} + a_2 \begin{pmatrix} S_{X,2}^2 \\ S_{Y,2}^2 \end{pmatrix} \tag{39}$$

where

$$n = n_1 + n_2, \quad a_1 = \frac{n_1 - 1}{n-2}, \quad a_2 = \frac{n_2 - 1}{n-2}$$

is also MLE for $(\sigma_X^2, \sigma_Y^2)^T$ respectively. It can be shown that

$$\begin{gathered} E[S_X^2] = \sigma_X^2,\ E[S_Y^2] = \sigma_Y^2, \\ \mathrm{Var}[S_X^2] \to \frac{2\sigma_X^4}{n},\ \mathrm{Var}[S_Y^2] \to \frac{2\sigma_Y^4}{n},\ \mathrm{Cov}(S_X^2, S_Y^2) \to \frac{2\rho^2\sigma_X^2\sigma_Y^2}{n} \end{gathered} \tag{40}$$

We have

$$\sqrt{n}\left[\begin{pmatrix} S_{X,1}^2 \\ S_{Y,1}^2 \end{pmatrix} - \begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\begin{pmatrix} \sigma_X^4 & \rho^2\sigma_X^2\sigma_Y^2 \\ \rho^2\sigma_X^2\sigma_Y^2 & \sigma_Y^4 \end{pmatrix}\right] \tag{41}$$

If we let $g(x) = x^{-\frac{1}{2}}$, and apply δ-method to equation (41), we obtain

$$\sqrt{n}\left[\begin{pmatrix} S_X^{-1} \\ S_Y^{-1} \end{pmatrix} - \begin{pmatrix} \sigma_X^{-1} \\ \sigma_Y^{-1} \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma_X^{-2} & \rho^2\sigma_X^{-1}\sigma_Y^{-1} \\ \rho^2\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\right] \tag{42}$$

It follows from equation (42) that $\mathrm{Corr}(S_X^{-1}, S_Y^{-1}) \to \rho^2$.

5 Conclusion

State the major findings
This article discusses the relationship between population correlation ρ and the corresponding test statistics correlation ρ_T. We investigate ρ_T for test statistics of the form $(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}, \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y})$ (see equation (7)), where the denominator

is the standard error of the numerator. Assuming independence between $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ and (S_X, S_Y), we derive the formula for test statistics correlation ρ_T, and show that ρ_T may not equal population correlation ρ.

In two group comparison setting, we conclude that $\rho_T = \rho$ when S_X (or S_Y) is constant with respect to $\boldsymbol{X}$ (or $\boldsymbol{Y}$). That is, $\rho_T = \rho$ under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is the case for two sample z-test. However, when S_X (or S_Y) is a function of $\boldsymbol{X}$ (or $\boldsymbol{Y}$), as is the case of two sample t-test, this equality may not hold. For two sample t-test, we prove that $\rho_T = \rho$ only if the null in equation (6) is true for all the tests considered, and that $|\rho_T| \leq |\rho|$ otherwise. In the case where one test is true null and the other true alternative, ρ_T is directly proportional to ρ, while when both tests are true alternatives, ρ_T is quadratic function of ρ.

State the practical meaningness of the findings

We note that cares need to be taken when estimating correlations between test statistics. In gene expression analysis, the two sample t-test [1, 5, 14] or moderated t-test [17] are used to calculate test statistics for DE detection, and the sample correlation (after treatment effects nullified) are used to adjust for correlation between those test statistics. Our study shows that, however, for DE genes, ρ_T may be overestimated when two genes are positively correlated, and underestimated when they are negatively correlated. If there are true DE genes whose expression levels are correlated in either way, the VIF may not be accurately estimated in [17], resulting in biased test for their enrichment analysis (REF our paper?). Our results also indicates that the variance of ρ_T may also be overestimated in [5], which leads to larger variation in estimating their conditional FDP.

Acknowledge the study's limitations

Theorem 1 and the subsequent results hold when the following two assumptions are met: 1) the test statistic has the of the form $\boldsymbol{a}^T\boldsymbol{X}/S_X$, and 2) $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are independent. In practice, both assumptions are vulnerable. The test statistic may take different forms, depending on many factors such as the nature of the data (RNA-Seq or microarray), the experimental design structure, and the statistical hypothesis to be tested. The independence assumption between $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are unlikely to hold unless the statistic is derived from two sample t-test for normally distributed random variables. Therefore, the application of Theorem 1 is very limited. Yet one goal of this study is to raise awareness that the equality of ρ_T and ρ should not be taken for granted. In the future, we will explore the relationship between ρ_T and ρ for more general cases and for other types of statistics.

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

[14] Qiu, X., Klebanov, L., and Yakovlev, A. (2005b). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

[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.