Sample correlation and test statistics correlation

1 Introduction

Overall theme: Why would we conduct this study? Or what is the motivation of this study?

large scale hypothesis testing in expression analysis

Gene expression analysis involves hypothesis testing for tens of thousands of genes simultaneously in biological research. One common feature of such testing is that a summary statistic is calculated for each gene to measure the magnitude of differential expression (DE). The test statistics are often of familiar form, coming from a simple two group comparison. The statistics themselves or their corresponding p-values are then pooled together, treated as known quantities, and used in downstream analysis such as false discovery rate (FDR) or enrichment analysis.

Correlation among test statistics need to be considered.

The downstream analysis, however, may be sensitive to the presence of correlations among gene expression level. Even modest such correlation can dangerously inflate the apparent FDR. For example, Efron (2007) showed in a simulation study that for a nominal FDR of 0.1, the actual FDR can easily vary by a factor of 10. Therefore, attempts have been made to account for the inter-gene correlation. A first approach is permutation or resampling of biological samples, which generates the null distributions of test statistics (or their associated p-values) with the inter-gene correlation preserved. The Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) falls into this category. A second approach works with the inter-gene correlation directly, by estimating the test statistics correlation. Efron (2007) modeled the distribution of the z-values correlation [explained in (1)] and estimated some dispersion variate A conditioning on which the False Discovery Proportion (FDP) was calculated. Wu and Smyth (2012) estimated the variance inflation factor (VIF) associated with inter-gene correlation and incorporated it into parametric and rank-based enrichment testing procedures. This paper concerns a nuisance aspect of the second approach, the relationship between sample correlations and test statistics correlations.

Key question: Are expression level correlations the same as test statistics correlation?

There's no way to obtain the correlation structure of test statistics without replicating the experiemnt. In the hypothesis testing procedures, only a single test statistic can be obtained for each gene. To incorporate inter-gene correlation into downstream analysis, the sample correlations, after nullifying gene treatment effects (we term as **residual sample correlation**), are used to replace the correlations of test statistics. The validity of such replacement is then demonstrated by small simulations. Efron (2007) used z-values converted from corresponding two sample t-test statistics by

$$z = \Phi^{-1}(G_0(t)) \tag{1}$$

where Φ is the cumulative distribution function (CDF) for $N(0,1)$ and G_0 is a putative null CDF for t-values. The correlation of z-values were approximated by residual sample correlation, since it was demonstrated via simulation that the distribution of residual sample correlation applies to that of z-value correlation. Barry et al. (2008) showed by Monte Carlo simulation of gene expression data that a nearly linear relationship holds between test statistic correlation and residual sample correlation under several standard experimental design. The sample correlation was then used to account for test statistics correlation in their bootstrap method of enrichment test. Wu and Smyth (2012) assumed that genewise t-test statistics correlation is the same as residual sample correlation, and calculated the mean of all pairwise sample correlation to estimate the VIF, a vital factor in adjusting for inter-gene

correlation in their enrichment test procedures. In all of the three works, it was shown by simulation only the equivalency of sample correlation coefficient and test statistics correlation coefficient, whether in distribution or numerically. It has, to the best of our knowledge, not yet been fully explored in the context of two group comparison.

Relevant but different work

A relevant research was done by Qiu et al. (2005), 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.

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 equivalance 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.

2 Methods

General setup

For simplicity, we consider only two genes with correlated expression values in a treatment/control comparison at a time. Suppose there are n samples each for treatment and control groups. Let gene 1 have expression levels $G_1 = (X_{11}, \ldots, X_{1n}, Y_{11}, \ldots, Y_{1n})$, and gene 2 $G_2 = (X_{21}, \ldots, X_{2n}, Y_{21}, \ldots, Y_{2n})$, where X and Y denote measurements from treatment and control, respectively. Suppose the expression values satisfy the following (A1-A3) assumptions:

A1): $\boldsymbol{X}_{j_1}$s and $\boldsymbol{Y}_{j_2}$s (j indexes sample) follow a bivariate normal distribution

$$\boldsymbol{X}_{j_1} = \begin{pmatrix} X_{1j_1} \\ X_{2j_1} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right] \stackrel{\text{def}}{=} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$\boldsymbol{Y}_j = (Y_{1j_2}, Y_{2j_2})^T \sim N(\boldsymbol{\mu} + \boldsymbol{\Delta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)^T$ for $i = 1, \ldots, n$.

A2): $\text{Cov}(\boldsymbol{X}_{j_1}, \boldsymbol{X}_{j_2}) = \text{Cov}(\boldsymbol{Y}_{j_1}, \boldsymbol{Y}_{j_2}) = \boldsymbol{0}$ for $j_1 \neq j_2$.

A3): $\text{Cov}(\boldsymbol{X}_{j_1}, \boldsymbol{Y}_{j_2}) = \boldsymbol{0}$ for all j_1, j_2.

Note that under this framework, A1) states that the mean vectors may be differet for $\boldsymbol{X}$ and $\boldsymbol{Y}$, but their variance structures are the same; for sample j_1 and j_2, the "true" correlation between gene 1 and gene 2 is $\text{Cor}(X_{1j_1}, X_{2j_1}) = \text{Cor}\ (Y_{1j_2}, Y_{2j_2}) = \rho$. A2) and A3) assumes that the samples are independet of each other.

In a typical expression analysis, we are interested in testing whether there is DE between treatment and control, statistically formularized as

$$H_{0i} : \Delta_i = 0 \text{ Versus } H_{1i} : \Delta_i \neq 0, \quad i = 1, 2. \tag{2}$$

Definition 1 *The **sample correlation** is*

$$r_{sample} = \frac{r_X + r_Y}{2} \tag{3}$$

where

$$r_X = \frac{\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)}{\sqrt{\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2}\sqrt{\sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2}} \tag{4}$$

$$r_Y = \frac{\sum_{j=1}^{n}(y_{1j} - \bar{y}_1)(y_{2j} - \bar{y}_2)}{\sqrt{\sum_{j=1}^{n}(y_{1j} - \bar{y}_1)^2}\sqrt{\sum_{j=1}^{n}(y_{2j} - \bar{y}_2)^2}} \tag{5}$$

and $\bar{X}_i = \sum_{j=1}^{n} X_{ij}/n$ *and* $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$ *for* $i = 1, 2$.
The pooled sample variance for gene i $(i = 1, 2)$ *is*

$$S_i^2 = \frac{(n-1)S_{X_i}^2 + (n-1)S_{Y_i}^2}{n-1+n-1} = \frac{S_{X_i}^2 + S_{Y_i}^2}{2} \tag{6}$$

where $S_{X_i}^2$ *and* $S_{Y_i}^2$ *are sample variances for treatment and control respectively.*

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 (1915).
For gene $i = 1, 2$, $\bar{X}_i \sim N(\mu_i, \sigma_i^2/n)$ and $\bar{Y}_i \sim N(\mu_i + \Delta_i, \sigma_i^2/n)$. Define

$$U_i = \bar{X}_i - \bar{Y}_i \tag{7}$$

The two sample z-test statistic is

$$Z_i = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{\frac{\sigma_i^2}{n} + \frac{\sigma_i^2}{n}}} = \frac{U_i}{\sqrt{2\sigma_i^2/n}} \sim N(\Delta_i, 1), \tag{8}$$

and the two sample t-test statistic for pooled variance is given by

$$T_i = \frac{\bar{X}_i - \bar{Y}_i}{S_i\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{U_i}{S_i\sqrt{\frac{2}{n}}} \sim t_{2n-2}(\Delta_i). \tag{9}$$

where S_i the pooled sample standard error defined in (6) and Δ_i is the non-center parameter. To show the relationship between test statistics correlation and true expression level correlation, we introduce the bivariate chi square distribution.

Lemma 2 *Let* $\boldsymbol{X}_j = (X_{1j}, X_{2j}), j = 1 \ldots, n$ *be independent random variables satisfying (A1), then* $\boldsymbol{W} = (W_{X_1}, W_{X_2}) = ((n-1)S_{X_1}^2/\sigma_1^2, (n-1)S_{X_2}^2/\sigma_2^2)$ *follows a* ***bivariate chi square distribution*** *with density (Joarder, 2009)*

$$f(w_1, w_2) = \frac{2^{-n}(w_1 w_2)^{(n-3)/2} e^{-\frac{w_1+w_2}{2(1-\rho^2)}}}{\sqrt{\pi}\Gamma(\frac{m}{2})(1-\rho^2)^{(n-1)/2}} \sum_{k=0}^{\infty}[1 + (-1)^k]\left(\frac{\rho\sqrt{w_1 w_2}}{1-\rho^2}\right)^k \frac{\Gamma(\frac{k+1}{2})}{k!\Gamma(\frac{k+n-1}{2})} \tag{10}$$

for $n - 1 > 2$ *and* $-1 < \rho < 1$.

The nice feature about Lemma 2 is that it is possible to separate the denominator from the numerator in calculating the expected product of T_1 and T_2 defined in (9), when T_1 and T_2 are correlatd.

Lemma 3 $\boldsymbol{U} = (U_1, U_2)$ *is independent of* $\boldsymbol{S} = (S_1, S_2)$, *where* $\boldsymbol{U}$ *and* S_i *are defined in (9) and (6) respectively.*

Proof: By lemma (2), the density function of $\boldsymbol{S}_X^2 = (S_{X_1}^2, S_{X_2}^2)$ only involves $\sigma_1^2, \sigma_2^2, \rho$ and sample size n, therefore we can denote its pdf $f(S_{X_1}^2, S_{X_2}^2)$ by some function $g(\sigma_1^2, \sigma_2^2, \rho, n)$. Note also that $\boldsymbol{S}^2 = \big((S_{X_1}^2 + S_{Y_1}^2)/2, (S_{X_2}^2 + S_{Y_2}^2)/2\big)$ is a linear combination of two independent bivariate chi square random variable $\boldsymbol{S}_X^2$ and $\boldsymbol{S}_Y^2$, its distribution can be expressed by another function $h(\sigma_1^2, \sigma_2^2, \rho, n)$. Therefore $\boldsymbol{S} = (S_1, S_2)$ is an ancillary statistic for $\boldsymbol{\Delta}$. On the other hand, it can be shown that $\boldsymbol{U} = (U_1, U_2)$ is a complete sufficient statistic for $\boldsymbol{\Delta}$. It follows by Basu's theorem that $\boldsymbol{U}$ and $\boldsymbol{S}$ are independent.

Following Lemma (3), U_1U_2 is also independent of $\frac{1}{S_1S_2}$, and therefore $E(\frac{U_1}{S_1} \cdot \frac{U_2}{S_2})$ can be expressed as $E(U_1U_2)E(\frac{1}{S_1S_2})$. Additionally, if we know $\mathrm{Cor}(\frac{1}{S_1}, \frac{1}{S_2})$, then the t-test statistics correlation can be accurately represented by Lemma (4).

Lemma 4 *Under (A1-A3), the two sample t-test statistics correlation can be expressed by*

$$Cor(T_1, T_2) = \frac{r_s \frac{\Delta_1\Delta_2}{\sigma_1\sigma_2} n(A-B) + 2\rho B + 2r_s\rho(A-B)}{\sqrt{\left[\frac{\Delta_1^2}{\sigma_1^2} n(A-B) + 2A\right]\left[\frac{\Delta_2^2}{\sigma_2^2} n(A-B) + 2A\right]}} \tag{11}$$

where

$$r_s = Cor(\frac{1}{S_1}, \frac{1}{S_2}) \tag{12}$$

$$A = \frac{n-1}{n-2}, \tag{13}$$

$$B = \frac{(n-1)\Gamma^2(n-3/2)}{\Gamma^2(n-1)}, \tag{14}$$

Proof: First note that

$$\begin{aligned}
\mathrm{Cov}(T_1, T_2) &= E(T_1T_2) - E(T_1)E(T_2) \\
&= E(c_0\frac{U_1}{S_1} \cdot c_0\frac{U_2}{S_2}) - E(c_0\frac{U_1}{S_1})E(c_0\frac{U_2}{S_2}) \\
&= c_0^2\left[E(U_1U_2)E(\frac{1}{S_1S_2}) - E(\frac{U_1}{S_1})E(\frac{U_2}{S_2})\right] \qquad \text{(by lemma 3)}
\end{aligned}$$

where $c_0 = \sqrt{\frac{n}{2}}$ and $\mathrm{Var}(T_1) = \mathrm{Var}(c_0\frac{U_1}{S_1}) = c_0^2\mathrm{Var}(\frac{U_1}{S_1})$. Note that

$$\mathrm{Cor}(T_1, T_2) = \frac{\mathrm{Cov}(T_1, T_2)}{\sqrt{\mathrm{Var}(T_1)\mathrm{Var}(T_2)}} = \frac{E(U_1U_2)E(\frac{1}{S_1S_2}) - E(\frac{U_1}{S_1})E(\frac{U_2}{S_2})}{\sqrt{\mathrm{Var}(\frac{U_1}{S_1})\mathrm{Var}(\frac{U_2}{S_2})}} \tag{15}$$

We need to calculate $E(U_1U_2)$, $E(\frac{1}{S_1S_2})$, $E(\frac{U_i}{S_i})$ and $\mathrm{Var}(\frac{U_i}{S_i})$ for $i = 1, 2$.

1. Note that $U_i \sim N(\Delta_i, \frac{2\sigma_i^2}{n}), i = 1, 2$.

$$E(U_1U_2) = \mathrm{Cov}(U_1, U_2) + E(U_1)E(U_2) = \rho\frac{2\sigma_1\sigma_2}{n} + \Delta_1\Delta_2 \tag{16}$$

2. Since $\frac{(n-1)S_{X_1}^2}{\sigma_1^2}$ and $\frac{(n-1)S_{Y_1}^2}{\sigma_1^2}$ are independent and follow $\chi^2(n-1)$, we have $W_{S_1} = \frac{2(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(2n-2)$. It can be shown that

$$E(W_{S_1}^k) = \frac{2^k\Gamma(n-1+k)}{\Gamma(n-1)}$$

Therefore

$$E\left(\frac{1}{S_1}\right) = \frac{\sqrt{n-1}\Gamma(n-\frac{3}{2})}{\sigma_1\Gamma(n-1)} = \frac{\sqrt{A}}{\sigma_1}, \quad \mathrm{Var}\left(\frac{1}{S_1}\right) = \frac{n-1}{\sigma_1^2}\left[\frac{1}{n-2} - \frac{\Gamma^2(n-\frac{3}{2})}{\Gamma^2(n-1)}\right] = \frac{A-B}{\sigma_1^2}$$

Note that $r_s = \text{Cor}(\frac{1}{S_1}, \frac{1}{S_2})$, we have

$$E(\frac{1}{S_1 S_2}) = E(\frac{1}{S_1})E(\frac{1}{S_2}) + r_s\sqrt{\text{Var}(\frac{1}{S_1})\text{Var}(\frac{1}{S_2})} = \frac{A}{\sigma_1\sigma_2} + r_s\frac{A-B}{\sigma_1\sigma_2} \tag{17}$$

3. $U_i \sim N(\Delta_i, \frac{2\sigma_i^2}{n})$ and $\frac{2(n-1)S_i^2}{\sigma_i^2} \sim \chi^2(2n-2)$ and by Lemma 3 they are independent, we have

$$\frac{\frac{U_i-\Delta_i}{\sqrt{2\sigma_i^2/n}}}{\frac{2(n-1)S_i^2}{\sigma_i^2}/(2n-2)} = \frac{U_i-\Delta_i}{S_i}\sqrt{\frac{n}{2}} \sim t(2n-2)$$

It follows from $E\left(\frac{U_i-\Delta_i}{S_i}\sqrt{\frac{n}{2}}\right) = 0$ and $\text{Var}\left(\frac{U_i-\Delta_i}{S_i}\sqrt{\frac{n}{2}}\right) = \frac{n-1}{n-2}$ that

$$E\left(\frac{U_i}{S_i}\right) = \frac{\Delta_i}{\sigma_i}\sqrt{A} \tag{18}$$

$$\text{Var}\left(\frac{U_i}{S_i}\right) = \frac{2}{n}A + \frac{\Delta_i^2}{\sigma_i^2}(A-B) \tag{19}$$

Finally, the test statistics correlation (11) is obtained by plugging (16–19) into (15).

Up to now we have obtained an exact expression for $\text{Cor}(T_1, T_2)$, which depends not only on the sample size n, but also on Δ/σ, the magnitude of DE. The rest of this section discusses asymptotic property of $\text{Cor}(T_1, T_2)$ for large sample size.

Lemma 5 *Let* $A = \frac{n-1}{n-2}$, $B = \frac{(n-1)\Gamma^2(n-3/2)}{\Gamma^2(n-1)}$, *then the following results hold:*

1. $\lim_{n\to\infty} A = 1$.
2. $\lim_{n\to\infty} B = 1$.
3. $\lim_{n\to\infty} n(A-B) = \frac{1}{4}$.

Proof: We will use second order Stirling's formula

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + \frac{1}{12n}) \tag{20}$$

By Stirling's formula (20) and $\Gamma(n+\frac{1}{2}) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$

$$\begin{aligned}
\frac{\Gamma^2(n-\frac{3}{2})}{\Gamma^2(n-1)} &= \left[\frac{\Gamma(n-2+\frac{1}{2})}{\Gamma(n-1)}\right]^2 \\
&= \left[\frac{(2n-4)!}{[(n-2)!]^2}\left(\frac{1}{4}\right)^{(n-2)}\right]^2 \pi \\
&\approx \left[\frac{\sqrt{2\pi(2n-4)}\left(\frac{2n-4}{e}\right)^{2n-4}\left(1+\frac{1}{12(2n-4)}\right)}{\left[\sqrt{2\pi(n-2)}\left(\frac{n-2}{e}\right)^{2n-4}\left(1+\frac{1}{12(n-2)}\right)\right]^2}\left(\frac{1}{2}\right)^{2n-4}\right]^2 \pi \\
&= \frac{(n-2)\left[(n-2)+\frac{1}{24}\right]^2}{(n-2+\frac{1}{12})^4}
\end{aligned}$$

therefore $\lim_{n\to\infty} A = \lim_{n\to\infty} B = 1$. Next,

$$\begin{aligned}\lim_{n\to\infty} n(A-B) &= \lim_{n\to\infty} n\left[\frac{n-1}{n-2} - \frac{(n-2)\left[(n-2)+\frac{1}{24}\right]^2}{(n-2+\frac{1}{12})^4}\right] \\ &= \lim_{n\to\infty} \frac{\frac{1}{4}n(n-1)(n-2)^3 + o\left((n-2)^4\right)}{(n-2)\left[(n-2)^4 + o\left((n-2)^4\right)\right]} \\ &= \frac{1}{4}\end{aligned}$$

Application of Lemma 5 to (11) gives the limit of test statistics correlation,

$$\rho(T_1, T_2) = \lim_{n\to\infty} \mathrm{Cor}(T_1, T_2) = \frac{\rho + \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2} r_{\mathrm{s}}}{\sqrt{\left[1+\frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1+\frac{\Delta_2^2}{8\sigma_2^2}\right]}} \tag{21}$$

where r_{s} is defined in (12). When $\boldsymbol{\Delta} = \mathbf{0}$ then $\lim_{n\to\infty} \mathrm{Cor}(T_1, T_2) = \rho$; but when $\boldsymbol{\Delta} \neq \mathbf{0}$, $\lim_{n\to\infty} \mathrm{Cor}(T_1, T_2) \neq \rho$ in general. In the next section, we will discuss about it in further detail.

3 Results

In section 2 we derived the exact expression of statistics correlation coefficient for two sample t-test. In the first part of this section, we conclude theoretically that test statisitcs correlation coefficient and sample correlation coefficient 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 statisitcs correlation $r_{\text{statistics}}$ may be very different from true correlation ρ or sample correlation r_{sample}.

3.1 Theory

Theorem 1 *Under (A1-A3), $Cor(Z_1, Z_2) = \rho$ for two sample z-test.*

Proof: Note that

$$\begin{aligned}\mathrm{Cov}(Z_1, Z_2) &= \frac{1}{\sqrt{\frac{2\sigma_1^2}{n}\frac{2\sigma_2^2}{n}}}\mathrm{Cov}(\bar{X}_1 - \bar{Y}_1, \bar{X}_2 - \bar{Y}_2) \\ &= \frac{n}{2\sigma_1\sigma_2}[\mathrm{Cov}(\bar{X}_1, \bar{X}_2) + \mathrm{Cov}(\bar{Y}_1, \bar{Y}_2)] \\ &= \frac{n}{2\sigma_1\sigma_2}[\mathrm{Cov}(\sum_{i=1}^n X_{1i}, \sum_{i=1}^n X_{2i})/n^2 + \mathrm{Cov}(\sum_{j=1}^n Y_{1j}, \sum_{j=1}^n Y_{2j})/n^2] \\ &= \frac{n}{2\sigma_1\sigma_2}[\sum_{i=1}^n \mathrm{Cov}(X_{1i}, X_{2i})/n^2 + \sum_{j=1}^n \mathrm{Cov}(Y_{1j}, Y_{2j})/n^2] \\ &= \frac{n}{2\sigma_1\sigma_2}[\frac{n\rho\sigma_1\sigma_2}{n^2} + \frac{n\rho\sigma_1\sigma_2}{n^2}] \\ &= \rho\end{aligned}$$

Therefore it follows that

$$\mathrm{Cor}(Z_1, Z_2) = \frac{\mathrm{Cov}(Z_1, Z_2)}{\sqrt{\mathrm{Var}(Z_1)}\sqrt{\mathrm{Var}(Z_2)}} = \rho \tag{22}$$

Theorem 1 states that in the context of z-test, the statistics correlation remains the same as expression level correlation, regardless of true null or true alternative. In fact, it can be shown that this conclusion remains true as long as Z_1 and Z_2 are linear combinations of $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ respectively (proof shown in appendix). However, the conclusion does not always hold when it comes to two sample t-test.

Theorem 2 *Under (A1-A3), if both tests are true null (i.e., $\Delta_i = 0, i = 1, 2$ in (2)) , $Cor(T_1, T_2) \to \rho$ as $n \to \infty$.*

Proof: If null is true for both test or $\mathbf{\Delta} = 0$, then (11) reduces to

$$\mathrm{Cor}(T_1, T_2) = \left[\frac{B}{A} + \frac{(A-B)r_s}{A}\right]\rho \tag{23}$$

$$= \left[r_s \cdot 1 + (1 - r_s)\frac{B}{A}\right]\rho \tag{24}$$

The term in the square bracket is a weighted average of 1 and $\frac{B}{A}$, with the latter converging to 1 as n grows to infinity. Therefore $\lim_{n\to\infty} \mathrm{Cor}(T_1, T_2) = \rho$. Table (??) shows the test statistics correlation [calculated by (23)] for known r_s with growing sample sizes.

Table 1: My caption

	n	3	5	10	50	100
	B/A	0.785	0.920	0.969	0.995	0.997
$\frac{\mathrm{Cor}(T_1,T_2)}{\rho}$	$r_s = 0.2$	0.828	0.936	0.975	0.996	0.998
	$r_s = 0.5$	0.893	0.960	0.985	0.997	0.999
	$r_s = 0.8$	0.957	0.984	0.994	0.999	0.999

Theorem 3 *If at least one of $H_{0i}, i = 1, 2$ in (2) is not true, then $Cor(T_1, T_2)$ does not converge to ρ in general.*

This conclusion is straightforward from (21). In the following simulation study, we demonstrated that depending on the true value of $\mathbf{\Delta}$ (DE or not DE, up-regulated or down-regulated if DE), $\mathrm{Cor}(T_1, T_2)$ might be far from ρ in different ways, discussed below.

We show via simulation [figure (??)] that for ρ growing from -1 to 1, r_s in (12) has a "U" shape whose minimum is located near $\rho = 0$, and

$$0 \le r_s \le |\rho| \quad \text{ONLY BASED ON SIMULATION} \tag{25}$$

(25) is useful in comparing $\rho(T_1, T_2)$ and ρ. For $\rho < 0$

1. if $\Delta_1\Delta_2 > 0$, then gene 1 and gene 2 are DE in the same direction (both up-regulated or both down-regulated), then

$$\rho(T_1, T_2) = \frac{\rho + \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2}r_s}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \frac{\rho}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho$$

2. if $\Delta_1\Delta_2 < 0$, then gene 1 and gene 2 are DE in different directions (one up-regulaed and the other down-regulated), then by $r_s < -\rho$,

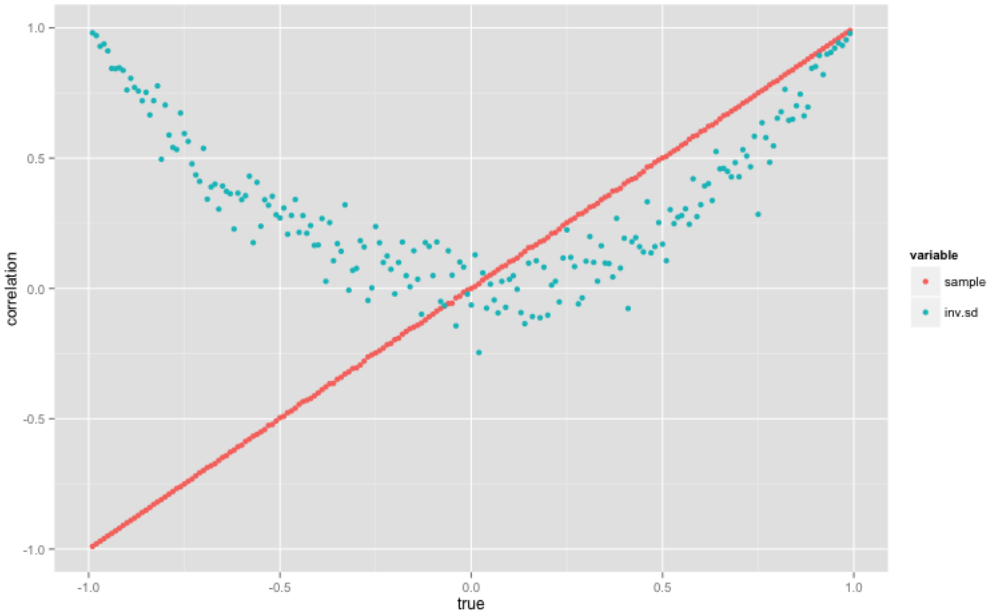
$$\rho(T_1, T_2) = \frac{\rho + \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2}r_s}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho\frac{1 - \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2}}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho$$

3. if $\Delta_1\Delta_2 = 0$, then one is DE but the other is not. Suppose gene 1 is not DE, then

$$\rho(T_1, T_2) = \frac{\rho}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]}} > \rho$$

Therefore in any case, $\rho(T_1, T_2) \ge \rho$ when $\rho < 0$. Similarly it can be shown that for $\rho > 0$, $\rho(T_1, T_2) \le \rho$.

Figure 1: $\text{Cor}(S_1^{-1}, S_2^{-1})$ against ρ



3.2 Simulation

The simulations are performed under three different testing procedures. The first two are based on normal distribution assumption, where we evaluate the true correlation and statistics correlation for z-test and t-test. In the third setting, we simulate correlated Poisson data to mimic RNA-Seq counts, and evaluate the relationship between the two for score test of Poisson regression.

For the normal case, we let

$$\begin{aligned} \boldsymbol{X}_i &= \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ -10 \end{pmatrix}, \begin{pmatrix} 0.1 & \rho\sqrt{0.1 \cdot 0.3} \\ \rho\sqrt{0.1 \cdot 0.3} & 0.3 \end{pmatrix} \right] \\ \boldsymbol{Y}_j &= \begin{pmatrix} Y_{1j} \\ Y_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 10 + \Delta_1 \\ -10 + \Delta_2 \end{pmatrix}, \begin{pmatrix} 0.1 & \rho\sqrt{0.1 \cdot 0.3} \\ \rho\sqrt{0.1 \cdot 0.3} & 0.3 \end{pmatrix} \right] \end{aligned} \tag{26}$$

with ρ growing continuously from -0.99 to 0.99 by 0.01. The sample size n is set to be 1000 ($j = 1, \ldots, 500$ for each group). For each given ρ, we generate $50,000$ samples for control group and another $50,000$ samples for the treatment group. The $50,000$ samples within each group are then randomly split into 100 blocks of size 500. Next, a pair is formed by taking one block (500 samples) from treatment and one blcok from control, mimicing one experiment for two group comparison. Therefore, 100 pairs are obtained to represent 100 replicates of the same experiment, from which 100 test statistics are computed for each gene.

The sample correlation r_{sample} is calculated as the average of sample correlation for each group by (4). The correlations between test statistics (z-statistics Z in 8 and t-statistics T in (9)) are calculated by the sample correlation of $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{100})$,

$$r_{\text{statistics}} = \frac{\sum_{j=1}^{100} (Z_{1j} - \bar{Z}_1)(Z_{2j} - \bar{Z}_2)}{\sqrt{\sum_{j=1}^{100} (Z_{1j} - \bar{Z}_1)^2 + \sum_{j=1}^{100} (Z_{2j} - \bar{Z}_2)^2}}, \text{ for } z\text{-test} \tag{27}$$

For t-test the correlation is calculated by (27), except the Z's being replaced by T. It is r_{sample}, $r_{\text{statistics}}$ and ρ that we are interested in comparing against. Specifically, we compared the three correlations for the following four cases:

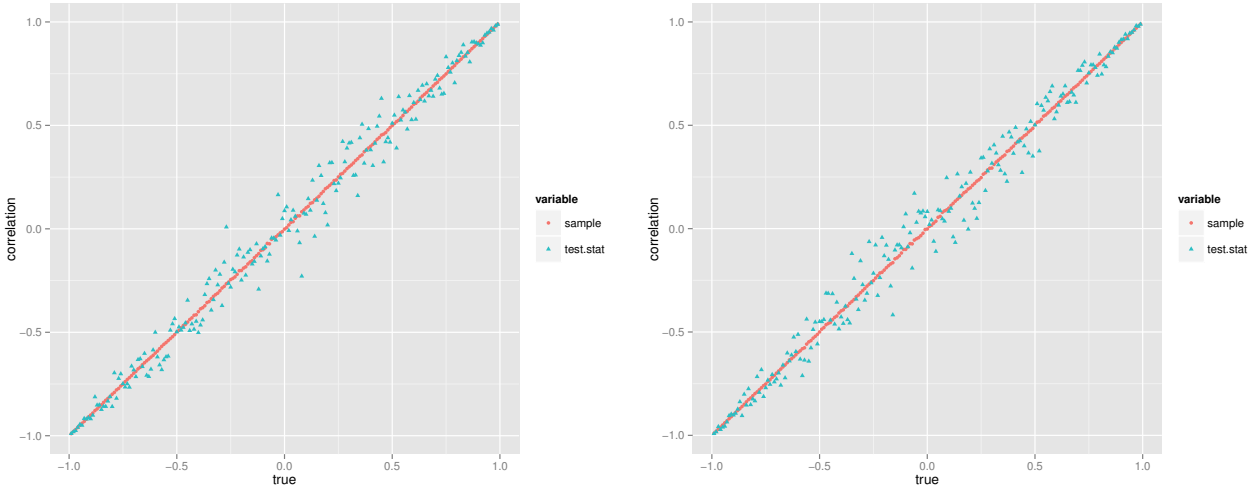
a) no DE genes;

b) DE in opposite directions;

c) DE in the same direction;

d) gene 1 DE and gene 2 null.

Figure (??) displays correlation for z-test statistics $r_{\text{statistics}}$, sample r_{sample} and true correlation ρ based on a)-d). The perfect positive dependence of the three quantities comes as no supprise for all situations.



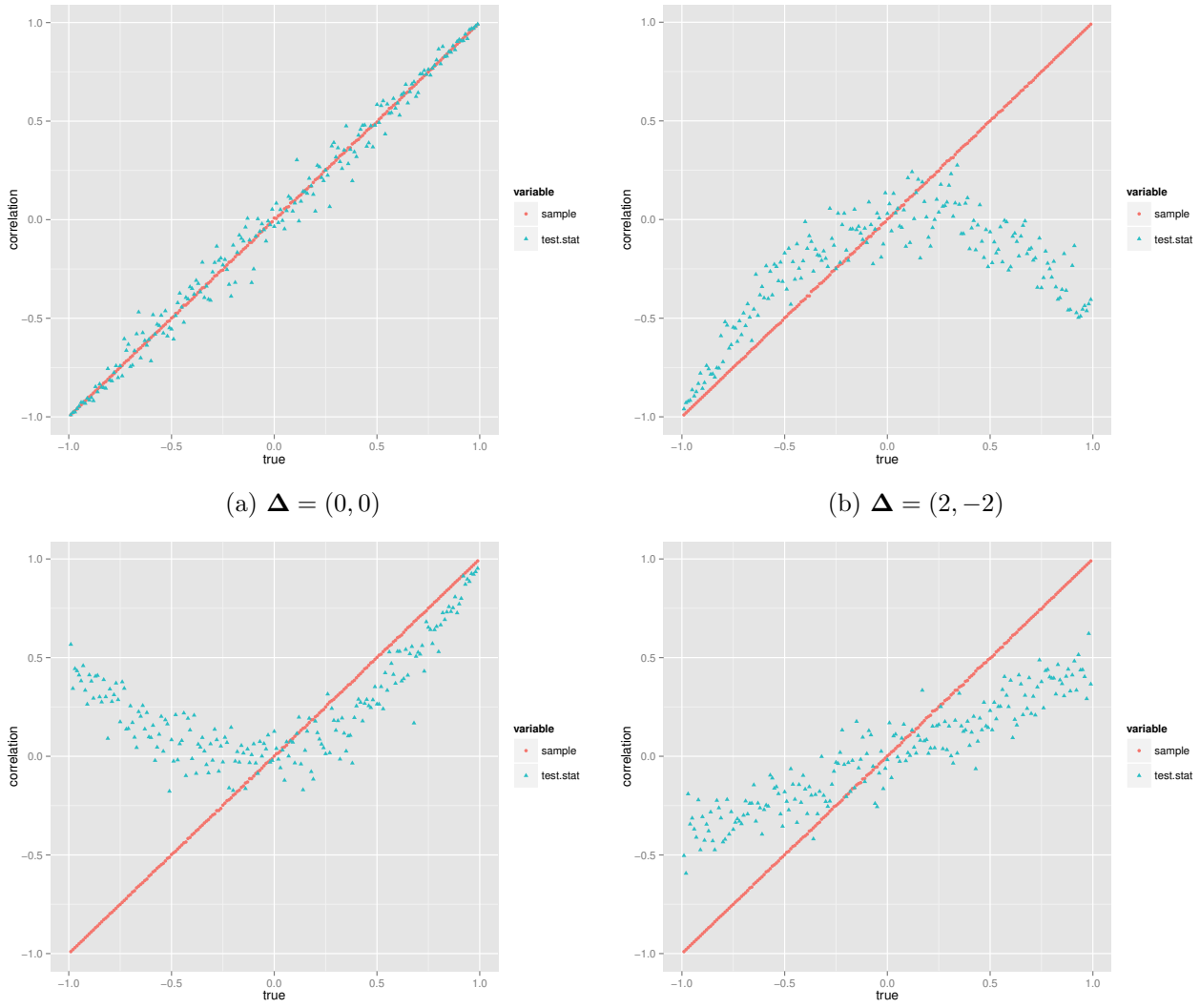
(a) $\boldsymbol{\Delta} = (0, 0)$ (b) $\boldsymbol{\Delta} = (2, -2)$

(c) $\boldsymbol{\Delta} = (2, 2)$ (d) $\boldsymbol{\Delta} = (-2, 0)$

Figure 2: Under z-test, the relationship between r_{sample} (solid dots), $r_{\text{statistics}}$ (triangles) and r_{true} (horizontal axis), for the case (a): gene 1 and gene 2 are not DE; (b): both genes are DE, but in different direction; (c): both genes are DE, in the same direction; (d): gene 1 is DE, but gene 2 is not. $\boldsymbol{\Delta}$ is the magnitude of DE.

For two sample t-test, the relationship between $r_{\text{statistics}}$ and ρ is more complicated. Figure (??) plots $r_{\text{statistics}}$ and r_{sample} against ρ. While the equivalence between those three holds when neither gene is DE [case a)], it fails as long as DE exists. $r_{\text{statistics}}$ is almost always negative, if genes are DE in different direction [case b)], and almost always positive if genes are DE in the same direction [case c)]. When only one gene is DE, $r_{\text{statistics}}$ is positively proportional to ρ [case d)]. Note that in all cases, $|r_{\text{statistics}}| \leq |\rho|$, in other words, the test statistics tend to be "less correlated" than the samples are.

Correlated Poisson data was simulated according to Madsen and Birkes (2013). Briefly, a 2-vector standard normal $\boldsymbol{Z}$ is first generated with correlation matrix $\boldsymbol{\Sigma}_Z$, and Z_i's are converted to $U_i = \Phi(Z_i)$



(a) $\boldsymbol{\Delta} = (0, 0)$ (b) $\boldsymbol{\Delta} = (2, -2)$

(c) $\boldsymbol{\Delta} = (2, 2)$ (d) $\boldsymbol{\Delta} = (-2, 0)$

Figure 3: Under t-test, the relationship between r_{sample} (solid dots), $r_{\text{statistics}}$ (triangles) and r_{true} (horizontal axis), for the case (a): gene 1 and gene 2 are not DE; (b): both genes are DE, but in different direction; (c): both genes are DE, in the same direction; (d): gene 1 is DE, but gene 2 is not. $\boldsymbol{\Delta}$ is the magnitude of DE.

where Φ is standard normal CDF. U_i's, uniform on (0, 1), are then trnasformed to $Y_i \equiv F_i^{-1}(U_i)$ with

$$F_i^{-1}(u) = \inf\{y : F_i(y) \geq u\} \tag{28}$$

The element in $\boldsymbol{\Sigma}_Z$ are chosen such that the desired Pearson correlation can be achieved. Technical details are available in Madsen and Birkes (2013) and thus not discussed here.

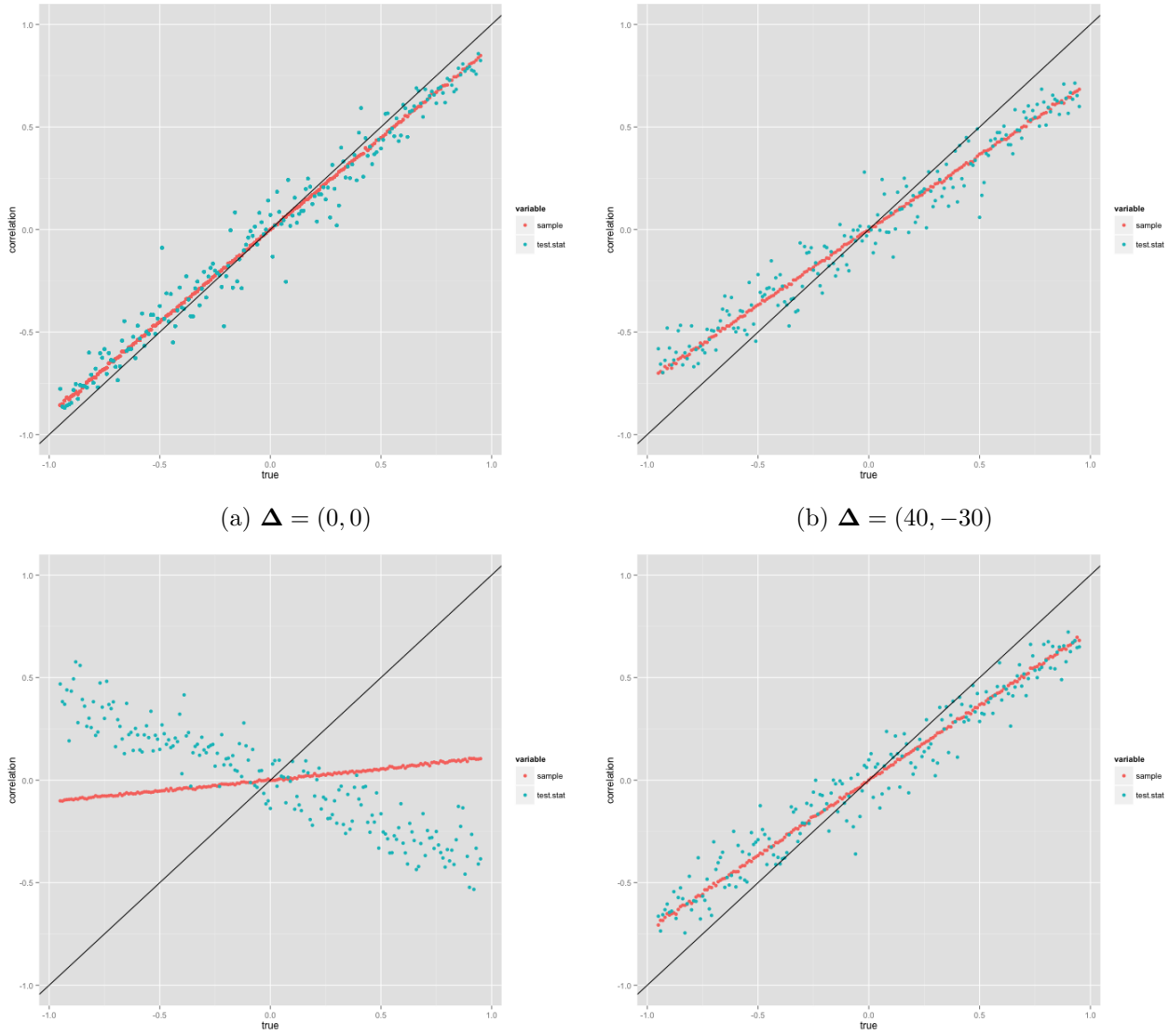
The two group comparison under Poisson regression are simulated as follows: for control group $X_{1i} \sim \text{Pois}(20)$ and $X_{2i} \sim \text{Pois}(50)$ with $\text{Cov}(X_{1i}, X_{2i}) = \rho$; for treatment group, a shift $\boldsymbol{\Delta}$ is added to the mean vectors, in other words, $Y_{1i} \sim \text{Pois}(20+\Delta_1)$ and $Y_{2i} \sim \text{Pois}(50+\Delta_2)$ with $\text{Cov}(Y_{1j}, Y_{2j}) = \rho$. The test statistics are calculated from score test (derivation is available in appendix),

$$U = \frac{\sqrt{\frac{n}{2}}(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y}_1 + \bar{y}_2}}. \tag{29}$$

Unlike the normal distribution whose shape is determined by both the mean and variance parameters, the shape of a Poisson distribution is totally determined by its mean parameter. For a score test statistic such as (29), the denominator and the numerator are no longer independent. Subsequently,

the derivation of test statistics correlation for t-test is invalid for Poisson regression. We will only demonstrate via simulation the relationship between $r_{\text{statistics}}$, r_{sample} and ρ.

Figure (??) presents the simulation under scenarios a)-d). The equivalence of $r_{\text{statistics}}$, r_{sample} and ρ still holds in general when neither gene is DE.



(a) $\boldsymbol{\Delta} = (0, 0)$ (b) $\boldsymbol{\Delta} = (40, -30)$

(c) $\boldsymbol{\Delta} = (40, 70)$ (d) $\boldsymbol{\Delta} = (40, 0)$

Figure 4: Under score test of Poisson regression, the relationship between r_{sample} (solid dots), $r_{\text{statistics}}$ (triangles) and r_{true} (horizontal axis), for the case (a): gene 1 and gene 2 are not DE; (b): both genes are DE, but in different direction; (c): both genes are DE, in the same direction; (d): gene 1 is DE, but gene 2 is not.

4 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 (2) 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 (Efron (2007), Barry et al. (2008)) or its moderated version (Wu and Smyth, 2012) 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 Wu and Smyth (2012), which may result in conservative test for enrichment analysis; the variance of $r_{\text{statistics}}$ may also be overestimated in Efron (2007), 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 (25) is true. Our simulation does suggest

$$r_s = \rho^2, \tag{30}$$

as shown in figure (**??**). If (30) 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 distrubutions, or for other hypothesis testing under different regression models (e.g., negative binomial regression).

5 Appendix

Test statistics correlation is the same as true correlation under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Let $\boldsymbol{A}$ be a nonzero vector of length $2n$, and $Z_1 = \boldsymbol{A}^T\boldsymbol{G}_1, Z_2 = \boldsymbol{A}^T\boldsymbol{G}_2$. Note that

$$\text{Cov}(Z_1, Z_2) = \boldsymbol{A}^T\text{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_2)\boldsymbol{A} = \boldsymbol{A}^T\text{diag}(\rho\sigma_1\sigma_2)\boldsymbol{A} = \rho\sigma_1\sigma_2\boldsymbol{A}^T\boldsymbol{A}.$$

and also

$$\text{Var}(\boldsymbol{A}^T\boldsymbol{G}_1) = \boldsymbol{A}^T\text{Var}(\boldsymbol{G}_1)\boldsymbol{A} = \sigma_1^2\boldsymbol{A}^T\boldsymbol{A}$$

It follows therefore

$$\text{Cor}(Z_1, Z_2) = \frac{\text{Cov}(Z_1, Z_2)}{\sqrt{\text{Var}(Z_1)\text{Var}(Z_2)}} = \frac{\rho\sigma_1\sigma_2\boldsymbol{A}^T\boldsymbol{A}}{\sigma_1\sigma_2\sqrt{\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{A}}} = \rho$$

Score test statistics correlation under Poisson regression

For a gene, let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the gene expression level, and $X = (1, \ldots, 1, 0, \ldots, 0)$ be the indicator of whether sample is from treatment or control group. A Poisson regression model

$$\begin{aligned} Y_i &\sim Pois(\mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

From the likelihood function

$$L = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}$$

we obtain the log-likelihood function

$$l(\beta_0, \beta_1) = \log L = \sum_{i=1}^{n} (y_i \log \mu_i - \log y_i! - \mu_i) \tag{31}$$
$$= \sum y_i(\beta_0 + \beta_1 x_i) - \sum \log y_i! - \sum \exp(\beta_0 + \beta_1 x_i)$$

For testing $H_0 : \beta_1 = 0$, the score test statistics is

$$U = [Z(\tilde{\beta})^T I^{-1}(\tilde{\beta}) Z(\tilde{\beta})]^{1/2}$$

where $\tilde{\beta} = (\hat{\beta}_0, 0)$. From (31) we have

$$\frac{\partial l}{\partial \beta_0} = \sum_i y_i - \sum_i \exp(\beta_0) \Rightarrow \quad \hat{\beta}_0 = \log(\bar{y})$$

Therefore

$$Z(\tilde{\beta}) = \begin{bmatrix} \sum_i y_i - \sum_i \exp(\beta_0 + \beta_1 x_i) \\ \sum_i y_i x_i - \sum_i \exp(\beta_0 + \beta_1 x_i) x_i \end{bmatrix} |_{\beta_1 = 0} = \begin{bmatrix} \sum y_i - \exp(\hat{\beta}_0) \\ \sum y_i x_i - \sum \exp(\hat{\beta}_0) x_i \end{bmatrix} = \begin{bmatrix} 0 \\ \sum y_i x_i - \bar{y} \sum x_i \end{bmatrix}$$

$$I(\tilde{\beta}) = \begin{bmatrix} \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i \\ \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i^2 \end{bmatrix} = \begin{bmatrix} \sum y_i & \bar{y} \sum x_i \\ \bar{y} \sum x_i & \bar{y} \sum x_i^2 \end{bmatrix}$$

and it follows that

$$U = [Z(\tilde{\beta})^T I^{-1}(\tilde{\beta}) Z(\tilde{\beta})]^{1/2} = \left(\frac{n(\sum y_i x_i - \bar{y} \sum x_i)^2}{\bar{y}[n \sum x_i^2 - (\sum x_i)^2]} \right)^{1/2} \tag{32}$$

To simplify the above expression, let's assume the first $n/2$ elements of $\boldsymbol{X}$ are 1, therefore we have $\sum x_i = \sum x_i^2 = n/2$

$$U = \sqrt{\frac{n(\sum_{i=1}^{n/2} y_i - \bar{y} \cdot n/2)^2}{\bar{y}[n \cdot n/2 - (n/2)^2]}} = \sqrt{\frac{\frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2}{\bar{y}_1 + \bar{y}_2}} = \pm \frac{\sqrt{\frac{n}{2}}(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y}_1 + \bar{y}_2}} \tag{33}$$

where $\bar{y}_1 = \frac{\sum_{i=1}^{n/2} y_i}{n/2}$ and $\bar{y}_2 = \frac{\sum_{i=n/2+1}^{n} y_i}{n/2}$ are just group means. It resembles a t test statistic.

References

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.

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

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

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

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

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.

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.

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.