Relation of Z-test statistics correlation and sample correlation

Under normal assumption. Let's assume gene 1 has expression level $G_1 = (X_{11}, \ldots, X_{1m}, Y_{11}, \ldots, Y_{1n})$ containing two treatments, similar for gene 2 $G_2 = (X_{21}, \ldots, X_{2m}, Y_{21}, \ldots, Y_{2n})$. Suppose in a general sense

$$X_{i1} \sim N(\mu_1, \sigma_1^2), \;\; Y_{j1} \sim N(\mu_2, \sigma_2^2), \quad X_{2i} \sim N(\mu_3, \sigma_3^2), \;\; Y_{2j} \sim N(\mu_4, \sigma_4^2) \tag{1}$$

Gene-Gene correlation

The correlation between two genes is defined, in my simulation, as

$$\rho = Cor(X_{1i}, X_{2i}) = Cor(Y_{1j}, Y_{2j}) \tag{2}$$

Sample correlation

Let $\bar{G}_1 = (\bar{X}_1, \bar{Y}_1)$ and $\bar{G}_2 = (\bar{X}_2, \bar{Y}_2)$, the sample correlation is defined as

$$\begin{aligned} Cor(G_1, G_2) &= \frac{Cov(G_1, G_2)}{\sqrt{Var(G_1)Var(G_2)}} \\ &= \frac{\sum_i (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) + \sum_j (Y_{1j} - \bar{Y}_1)(Y_{2j} - \bar{Y}_2)}{\sqrt{\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_j (Y_{1j} - \bar{Y}_1)^2}\sqrt{\sum_i (X_{2i} - \bar{X}_2)^2 + \sum_j (Y_{2j} - \bar{Y}_2)^2}} \end{aligned} \tag{3}$$

Z test statistics

Now the test statistics are

$$Z_1 = \frac{\bar{X}_1 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(\mu_1 - \mu_2, 1)$$

$$Z_2 = \frac{\bar{X}_2 - \bar{Y}_2}{\sqrt{\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n}}} \sim N(\mu_3 - \mu_4, 1)$$

therefore assuming samples are independent of each other

$$\begin{aligned} Cov(Z_1, Z_2) &= \frac{1}{c_0} Cov(\bar{X}_1 - \bar{Y}_1, \bar{X}_2 - \bar{Y}_2) \\ &= \frac{1}{c_0}[Cov(\bar{X}_1, \bar{X}_2) + Cov(\bar{Y}_1, \bar{Y}_2)] \\ &= \frac{1}{c_0}[Cov(\sum_{i=1}^{m} X_{1i}, \sum_{i=1}^{m} X_{2i})/m^2 + Cov(\sum_{j=1}^{n} Y_{1j}, \sum_{j=1}^{n} Y_{2j})/n^2] \\ &= \frac{1}{c_0}[\sum_{i=1}^{m} Cov(X_{1i}, X_{2i})/m^2 + \sum_{j=1}^{n} Cov(Y_{1j}, Y_{2j})/n^2] \\ &= \frac{1}{c_0}[\frac{\rho\sigma_1\sigma_3}{m} + \frac{\rho\sigma_2\sigma_4}{n}] \end{aligned}$$

where $c_0 = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\left(\frac{\sigma_3^2}{m} + \frac{\sigma_2^4}{n}\right)$. Note that $Var(Z_1) = 1$ we have

$$\rho(Z_1, Z_2) = \frac{Cov(Z_1, Z_2)}{\sqrt{Var(Z_1)}\sqrt{Var(Z_2)}} = \rho \cdot \frac{\frac{\sigma_1\sigma_3}{m} + \frac{\sigma_2\sigma_4}{n}}{\sqrt{(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})(\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n})}} \tag{4}$$

This equals to ρ if and only if

$$\frac{\frac{\sigma_1\sigma_3}{m} + \frac{\sigma_2\sigma_4}{n}}{\sqrt{(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})(\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n})}} = 1 \Rightarrow \sigma_1\sigma_4 = \sigma_2\sigma_3 \tag{5}$$

In a typical gene expression analysis, it is assumed that for the same gene, variance across different treatments are constant. Therefore $\sigma_1 = \sigma_2$ and $\sigma_3 = \sigma_4$, and the gene correlation and test statistic correlation are the same for Z-test.

Relation of T-test statistics correlation and sample correlation

Similarly, under assumption (12), but we don't know the $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$. For now, we assume within the same gene, the expression levels have the same variance (i.e., $\sigma_1^2 = \sigma_2^2, \sigma_3^2, = \sigma_4^2$,).

Pooled variance

The T test statistics (pooled variance) for gene 1 and gene 2 are

$$T_1 = \frac{\bar{X}_1 - \bar{Y}_1}{S_1\sqrt{\frac{1}{m} + \frac{1}{n}}}, T_2 = \frac{\bar{X}_2 - \bar{Y}_2}{S_2\sqrt{\frac{1}{m} + \frac{1}{n}}} \tag{6}$$

where

$$S_1^2 = \frac{(m-1)S_{X_1}^2 + (n-1)S_{Y_1}^2}{m+n-2}, \quad S_2^2 = \frac{(m-1)S_{X_2}^2 + (n-1)S_{Y_2}^2}{m+n-2}$$

Since $S_{X_1}^2, S_{Y_1}^2, S_{X_2}^2, S_{Y_2}^2$ are consistent estimators of $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ respectively. Therefore

$$S_1^2 \xrightarrow{d} \frac{(m-1)\sigma_1^2 + (n-1)\sigma_2^2}{m+n-2} \stackrel{def}{=} \delta_1^2, \qquad S_2^2 \xrightarrow{d} \frac{(m-1)\sigma_3^2 + (n-1)\sigma_4^2}{m+n-2} \stackrel{def}{=} \delta_2^2,$$

Therefore we have

$$\begin{aligned} Cov(T_1, T_2) &\approx \frac{1}{\delta_1\delta_2(\frac{1}{m} + \frac{1}{n})} Cov(\bar{X}_1 - \bar{Y}_1, \bar{X}_2 - \bar{Y}_2) \\ &= \frac{1}{\delta_1\delta_2(\frac{1}{m} + \frac{1}{n})}\left[\frac{\rho\sigma_1\sigma_3}{m} + \frac{\rho\sigma_2\sigma_4}{n}\right] \end{aligned}$$

Note that

$$Var(T_1) \approx \frac{1}{\delta_1^2(\frac{1}{m} + \frac{1}{n})} Var(\bar{X}_1 - \bar{Y}_1) = \frac{1}{\delta_1^2(\frac{1}{m} + \frac{1}{n})}(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$$

$$Var(T_2) \approx \frac{1}{\delta_1^2(\frac{1}{m}+\frac{1}{n})} Var(\bar{X}_2 - \bar{Y}_2) = \frac{1}{\delta_2^2(\frac{1}{m}+\frac{1}{n})}(\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n})$$

Therefore

$$\rho(T_1, T_2) = \frac{Cov(T_1, T_2)}{\sqrt{Var(T_1)Var(T_2)}} \approx \rho \cdot \frac{\frac{\sigma_1\sigma_3}{m} + \frac{\sigma_2\sigma_4}{n}}{\sqrt{(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})(\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n})}} \tag{7}$$

It resembles (4).

Unequal variances

The T test statistics in this case are

$$T_1 = \frac{\bar{X}_1 - \bar{Y}_1}{S_1}, T_2 = \frac{\bar{X}_2 - \bar{Y}_2}{S_2}$$

where

$$S_1^2 = \frac{S_{X_1}^2}{m} + \frac{S_{Y_1}^2}{n}, \;\; S_2^2 = \frac{S_{X_2}^2}{m} + \frac{S_{Y_2}^2}{n}$$

and we have

$$S_1^2 \xrightarrow{d} \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \overset{def}{=} \delta_1^2, \qquad S_2^2 \xrightarrow{d} \frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n} \overset{def}{=} \delta_2^2,$$

The correlation of T_1 and T_2 can be calculated by

$$Cov(T_1, T_2) \approx \frac{1}{\delta_1\delta_2}\left[\frac{\rho\sigma_1\sigma_3}{m} + \frac{\rho\sigma_2\sigma_4}{n}\right]$$

$$Var(T_1) \approx \frac{1}{\delta_1^2}(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}), \quad Var(T_2) \approx \frac{1}{\delta_2^2}(\frac{\sigma_3^2}{m} + \frac{\sigma_4^2}{n})$$

expression (7) holds for unequal variance case, too.

Sample correlation and Gene correlation

We have established the equality of inter-gene correlation and test statistic correlation (for both Z and T tests). However, it is interesting to point out that sample correlation does not mean inter-gene correlation.

An estimate of inter-gene correlation

$$\hat{\rho} = \frac{\sum_{i=1}^m (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^m (x_{i1} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^m (x_{i2} - \bar{x}_2)^2}}$$

or

$$\hat{\rho} = \frac{\sum_{j=1}^n (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sqrt{\sum_{j=1}^n (y_{i1} - \bar{y}_1)^2}\sqrt{\sum_{j=1}^n (y_{i2} - \bar{y}_2)^2}}$$

However, if we define the sample correlation as (3), that is,

$$Cor(G_1, G_2) = \frac{Cov(G_1, G_2)}{\sqrt{Var(G_1)Var(G_2)}}$$
$$= \frac{\sum_i (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) + \sum_j (Y_{1j} - \bar{Y}_1)(Y_{2j} - \bar{Y}_2)}{\sqrt{\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_j (Y_{1j} - \bar{Y}_1)^2}\sqrt{\sum_i (X_{2i} - \bar{X}_2)^2 + \sum_j (Y_{2j} - \bar{Y}_2)^2}}$$

NOTE: as long as (5) holds

$$Cor(Z_1, Z_2) \approx Cor(G_1, G_2)$$

where $\bar{G}_1 = (\bar{X}_1, \bar{Y}_1)$ and $\bar{G}_2 = (\bar{X}_2, \bar{Y}_2)$.

DE or NOT DE matters

If we look at the z-test statistic and t-test statistic,

$$Z = \frac{\bar{X}_1 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}, \quad T = \frac{\bar{X}_1 - \bar{Y}_1}{\sqrt{\frac{S_{X_1}^2}{m} + \frac{S_{Y_1}^2}{n}}}$$

Although we already established that $S_{X_1}^2 \xrightarrow{d} \sigma_1^2$ and $S_{Y_1}^2 \xrightarrow{d} \sigma_2^2$, the correlation of T statistics will not remain the same as correlation of Z statistics. Two reasons for that:

1. if m and n are small, then $S_{X_1}^2$ and $S_{Y_1}^2$ cannot be accurately estimated
2. if m and n are large, then the denominator will be very small, in which case a slight difference between the sample variance $S_{X_1}^2$ and true variance σ_1^2 will augment the difference of test statistics.

General Conclusion

The way we define the sample correlations really matters!

If we define the sample correlation as $\rho = \frac{1}{2}(\rho_1 + \rho_2)$ where $\rho_1 = Cor(X_1, X_2), \rho_2 = Cor(Y_1, Y_2)$, then sample correlation equals to test statistics correlation as long as (5) holds. We do care about whether there is DE or not (simulation study shows that).

For Z test, the correlation between Z statistics and correlation between expression value generally match, since within a gene, we assume they have the same variance across two treatments (i.e., $\sigma_1^2 = \sigma_2^2, \sigma_3^2, = \sigma_4^2,$).

However, for a typical T-test, the mean difference between the comparison will play a role. Denote $Y'_{1i} = Y_{1i} - \bar{Y}_1$ and similar for X_1, X_2 and Y_2. Unless the difference between Y'_i and Y_i is negligible (No DE), the sample correlation we obtained is different from T correlation. But if we dont remove the treatment mean, the correlation we obtained does not reflect the true gene correlation.

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

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

The likelihood function

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

And the log-likelihood function

$$\begin{aligned} l(\beta_0, \beta_1) = \log L &= \sum_{i=1}^{n} (y_i \log \mu_i - \log y_i! - \mu_i) \\ &= \sum y_i(\beta_0 + \beta_1 x_i) - \sum \log y_i! - \sum \exp(\beta_0 + \beta_1 x_i) \end{aligned} \tag{8}$$

Wald test

The first derivative of (8) with respect to β_0, β_1

$$\frac{\partial l}{\partial \beta_1} = \sum y_i x_i - \sum \exp(\beta_0 + \beta_1 x_i) x_i$$

$$\frac{\partial l}{\partial \beta_0} = \sum y_i - \sum \exp(\beta_0 + \beta_1 x_i)$$

The fisher information

$$I(\beta_0, \beta_1) = -E \begin{bmatrix} \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} \end{bmatrix}$$

Then the observed fisher information

$$\hat{I}(\beta_0, \beta_1) = \begin{bmatrix} \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i \\ \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i^2 \end{bmatrix}$$

The Wald statistics for $H_0 : \beta_1 = 0$ is therefore

$$T = \frac{\partial l}{\partial \boldsymbol{\beta}}^T [\hat{I}(\beta_0, \beta_1)]^{-1} \frac{\partial l}{\partial \boldsymbol{\beta}}$$

Since there is no analytical solution for β_1, we don't know what the relation of T and Y is.

Score test

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

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

In this case, $\tilde{\beta} = (\beta_0, 0)$. From (8) 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

$$U = [Z(\tilde{\beta})^T I^{-1}(\tilde{\beta}) Z(\tilde{\beta})]^{1/2} = \left(\frac{n(\sum y_i x_i - \bar{y} \sum x_i)^2}{\bar{y}[n \sum x_i^2 - (\sum x_i)^2]} \right)^{1/2} \tag{9}$$

To make it simpler, let's assume the first $n/2$ elements of $\boldsymbol{X}$ are 1 and $\sum_{i=1}^{n/2} y_i \geq \sum_{i=n/2+1}^{n} y_i$, therefore we have $\sum x_i = \sum x_i^2 = n/2$

$$U = \sqrt{\frac{n(\sum_{i=1}^{n/2} y_i - \bar{y} \cdot n/2)^2}{\bar{y}[n \cdot n/2 - (n/2)^2]}} = \sqrt{\frac{\frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2}{\bar{y}_1 + \bar{y}_2}} = \frac{\sqrt{\frac{n}{2}}(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y}_1 + \bar{y}_2}} \tag{10}$$

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. This resembles a t test statistics.

Simulation

Generate correlated Poisson random variables. Let $X_0 \sim Pois(\lambda_0), X_1 \sim Pois(\lambda_1), X_2 \sim Pois(\lambda_2)$ and X_0, X_1, X_2 are mutually independent of each other. Let $Y_1 = X_1 + X_0$, $Y_2 = X_2 + X_0$, then Y_1 and Y_2 are correlated.

$$Cov(Y_1, Y_2) = Cov(X_1 + X_0, X_2 + X_0) = Var(X_0) = \lambda_0$$

The correlation between Y_1 and Y_2 can then be expressed by

$$\rho(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1) Var(Y_2)}} = \frac{\lambda_0}{\sqrt{(\lambda_1 + \lambda_0)(\lambda_2 + \lambda_0)}}$$

Particularly, if we let $\lambda_1 = \lambda_2$, then $\rho = \frac{\lambda_0}{\lambda_0 + \lambda_1}$. More generally, $X_1 \sim Pois((1-\rho)\lambda), X_0 \sim Pois(\rho\lambda), X_2 \sim Pois((1-\rho)\lambda)$, then $Var(Y_1) = Var(Y_2) = \lambda$.

Note: there is an upper bound for this correlation

$$\rho_{Y_1, Y_2} = \frac{Cov(Y_1, Y_2)}{\sigma_{Y_1} \sigma_{Y_2}} = \frac{\sigma_X^2}{\sigma_{Y_1} \sigma_{Y_2}} \leq \frac{\min(\sigma_{Y_1}^2, \sigma_{Y_2}^2)}{\sigma_{Y_1} \sigma_{Y_2}} = \min\left(\frac{\sigma_{Y_1}}{\sigma_{Y_2}}, \frac{\sigma_{Y_2}}{\sigma_{Y_1}} \right)$$

For each pair of (ρ, λ), evaluate the sample correlation and score test statistics correlation UNDER THE NULL.

Negative Binomial regression

Simulation

Generate correlated negative binomial random numbers.

Lemma Let X_0, X_1, X_2 are i.i.d. negative binomial random variables, $X_i \sim NB(r_i, p)$ for $i = 0, 1, 2$. Then $Y = \sum_i X_i \sim NB(\sum r_i, p)$.

This parameterization yields if $X \sim NB(r, p)$

$$E[X] = \frac{r(1-p)}{p}$$

$$Var[X] = \frac{r(1-p)}{p^2}$$

equation of mean-dispersion parameterization gives

$$\frac{r(1-p)}{p} = \mu, \quad \frac{r(1-p)}{p^2} = \mu + k\mu^2$$

where μ and k are means and dispersion. Therefore

$$p = \frac{1}{1+k\mu}, \quad r = \frac{1}{k} \tag{11}$$

Let $Y_1 = X_0 + X_1$ and $Y_2 = X_0 + X_2$, then $Y_1 \sim NB(r_0 + r_1, p), Y_2 \sim NB(r_0 + r_2, p)$. We have

$$Cov(Y_1, Y_2) = Var(X_0) = \frac{r_0(1-p)}{p^2}$$

and the correlation of Y_1 and Y_2

$$\rho_{Y_1Y_2} = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}} = \frac{r_0}{\sqrt{(r_0 + r_1)(r_0 + r_2)}}$$

Particularly, if we let $r_1 = r_2$, then $\rho = \frac{r_0}{r_1 + r_0}$.

In simulation study, we let $X_0 \sim NB(\rho r, p)$, $X_1 \sim NB((1-\rho)r, p)$ and $X_2 \sim NB((1-\rho)r, p)$ and it follows that $Y_1 = X_0 + X_1 \sim NB(r, p)$, $Y_2 = X_0 + X_2 \sim NB(r, p)$, and $Cor(Y_1, Y_2) = \rho$.

For the null case, where there is no difference between treatment expression value $Y_{11}, \cdots, Y_{1,n/2}$ and control expression level $Y_{1,(n/2+1)}, \cdots, Y_{1n}$. From (11) we can see that μ should be a constant. Since my derivation of test statistic is under the assumption that dispersion is a constant for treatment/control, we also require k thus r to be constant. (If we want to simulate DE cases, simply put $\mu_1 \neq \mu_2$ or $p_1 \neq p_2$.)

For two genes, because of the way we simulate correlated NB data, it is required that p remain constant within treatment (or control), which means the term $k\mu$ is constant. Note since Y_1, Y_2 are

identically distributed (i.e. r is the same for two genes), both k and μ are the same for two genes.

To sum it up, we need k, (μ_1, μ_2) and the desired correlation ρ to the read count matrix under NB assumption.

Sample correlation and Two sample T-test correlation

Conclusion: the correlation of t test statistics converges to sample correlation if genes are NOT DE, and will generally be smaller than sample correlation if some genes are DE.

Under normal assumption. Let's assume gene 1 has expression level $G_1 = (X_{11}, \ldots, X_{1n}, Y_{11}, \ldots, Y_{1n})$ containing two treatments, similar for gene 2 $G_2 = (X_{21}, \ldots, X_{2n}, Y_{21}, \ldots, Y_{2n})$. Suppose in a general sense

$$X_{i1} \sim N(\mu_1, \sigma_1^2),\ \ Y_{j1} \sim N(\mu_1 + \Delta_1, \sigma_1^2),\quad X_{2i} \sim N(\mu_2, \sigma_2^2),\ \ Y_{2j} \sim N(\mu_2 + \Delta_2, \sigma_2^2) \tag{12}$$

Gene-Gene correlation

For the same gene, the samples are independent of each other. The correlation between two genes is defined, in my simulation, as

$$\rho = Cor(X_{1i}, X_{2i}) = Cor(Y_{1j}, Y_{2j}) \tag{13}$$

Then $U_1 = \bar{X}_1 - \bar{Y}_1 \sim N(\Delta_1, \frac{2\sigma_1^2}{n}), U_2 = \bar{X}_2 - \bar{Y}_2 \sim N(\Delta_2, \frac{2\sigma_2^2}{n})$ It follows that

$$\boldsymbol{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \Delta_1 \\ \Delta_2 \end{pmatrix}, \frac{2}{n} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

and define

$$S_{X_1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \Rightarrow \frac{(n-1)S_{X_1}^2}{\sigma_1^2} \sim \chi^2(n-1)$$

$$S_{Y_1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2 \Rightarrow \frac{(n-1)S_{Y_1}^2}{\sigma_1^2} \sim \chi^2(n-1)$$

$$S_1^2 = \frac{(n-1)S_{X_1}^2 + (n-1)S_{Y_1}^2}{n-1+n-1} = \frac{S_{X_1}^2 + S_{Y_1}^2}{2}$$

$$\boldsymbol{W}_X = \begin{pmatrix} W_{X_1} \\ W_{X_2} \end{pmatrix} = \begin{pmatrix} (n-1)S_{X_1}^2/\sigma_1^2 \\ (n-1)S_{X_2}^2/\sigma_2^2 \end{pmatrix} \sim \text{bivariate } \chi^2 \text{ distribution}$$

therefore the distribution of $W = \frac{W_X + W_Y}{2}$ can be derived from bivariate χ^2 distribution, and it's not related to the mean parameter $(\mu_1, \mu_2, \Delta_1, \Delta_2)$. By Basu's theorem, $\boldsymbol{U}$ is complete sufficient for mean, and W is ancillary for mean, then $\boldsymbol{U}$ is independent of $\boldsymbol{W}$. Subsequently (U_1, U_2) is independent of (S_1, S_2).