Journal of Data Science 18 (4), 632–648
October 2020
DOI: 10.6339/JDS.202010_18(4).0003

Gene Set Enrichment Analysis in RNA-Seq Data

Chen-An Tsai*[1] and Pei-Hsun Li[1]

[1]*Department of Agronomy, Division of Biometry, National Taiwan University, Taipei, Taiwan*

Abstract

To date, many gene set analysis (GSA) approaches have been developed for identifying differentially expressed gene sets using microarray data. However, these methods are not directly applicable to RNA-Seq data due to intrinsic difference between two data structures. When testing the differential expression of gene sets, there is a critical assumption that the members in each gene set are sampled independently in most GSA methods. It means that the genes within a gene set don't share a common biological function. The aim of this paper is twofold. First, we propose a powerful yet simple extension to GSA methods based on the de-correlation (DECO) algorithm that properly remove the correlation bias in the expression of each gene set. We then study the performance of our proposed method compared with other GSA methods through a real RNA-Seq dataset and simulation studies under various scenarios combining with four commonly used normalization methods. Second, we discuss the effect of the complex correlation structure of gene sets on four normalization methods. As a result, we found that our proposed method outperforms the others in terms of Type I error rate and empirical power. A comparative study on a public data showed that gene sets identified by our proposed method have better concordance with biological confirmed pathways than other methods.

Keywords *correlation bias; DECO; gene set analysis; RNA-Seq*

1 Introduction

Gene Set Analysis (GSA) originates from analyzing gene expression data. Up to the present, many methodologies for testing differential expression of gene sets are aggregately named Gene Set Analysis (GSA) approaches. According to Tian et al. (2005), GSA approaches can be grouped into two categories depending on the null hypothesis tests. One is that the genes in a gene set have the same level of association with the phenotype compared with the genes in the rest of the gene set. The other is that none of the genes in a gene set is associated with the phenotype of interest. The former called competitive approaches and the latter called self-contained approaches. Here, we focus on the latter group, which relies on permuting sample labels to generate null distributions. Also, it is known as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), which provides an insight into gene expression data.

RNA-Seq, also called whole transcriptome shotgun sequencing, is a novel technology, which uses next-generation sequencing (NGS) to reveal transcriptome profiling. In the past few years, RNA-Seq has almost taken the place of microarrays (Wang et al., 2009). There are many applications of RNA-Seq, including gene expression profiling, rare transcripts, SNPs coding, alternative splice sites, identification of novel transcripts, and so forth. Among these applications, the most popular use of RNA-Seq is to identify differentially expressed genes. In general, gene counts are modeled by using Poisson or Negative Binomial (NB) distribution. In Poisson distribution, the

*Corresponding author. Email: catsai@ntu.edu.tw.

Received August, 2019; Accepted April, 2020

mean is equal to the dispersion parameter. If the genes of each sample have identical expression level, then we can use Poisson distribution to model counts data. However, the Poisson distribution is not suitable for the over-dispersion in real data. To deal with the issue of over-dispersion, a common solution is to use the negative binomial distribution, which is known as the Gamma-Poisson mixture model. Due to the power of GSEA, many gene set analysis methods and tools for RNA-Seq have been developed recently (Ren et al., 2017; Wang and Cairns, 2013). However, most methods are based on algorithms developed for microarray data and challenges exist in applying the existing algorithms to RNA-Seq data. The differing correlation structure of enriched gene set may cause bias in gene set analysis.

In this paper, we propose an algorithm combining De-correlation (DECO) (Nam, 2010) with the sum of square t-statistic to identify gene sets with differential expression profiles between two biological conditions. Our algorithm offers an easy to implement GSA approach to quantifying underlying phenotypic differences with the inter-gene correlation embedded in a gene set, and allows us to rank the differential expression of gene sets. Since it was suggested that a normalized procedure should be applied to RNA-Seq counts data before GSA, we compare the performance of our proposed method to the existing multivariate tests in terms of the Type I error rate and the power under four different normalization methods. Among all simulation settings, our proposed method outperforms other methods in identifying differentially expressed gene sets. Among these normalizations, the Trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010) provides better estimates. In addition, we found that methods combined with DECO are more powerful than those without DECO when evaluating the power. Our simulation studies show that our proposed algorithm improves the accuracy of estimations. A real RNA-Seq data from lymphoblastoid cell lines of 69 unrelated Nigerian individuals is used to resent the performance of the proposed methods.

2 Materials and Methods

Recent advances of GSA provide a complementary and powerful approach to microarray technologies for global expression profile of biological systems. Methods for gene set analysis have successively revealed the underlying pathway activity variation associated with experimental conditions of interest. To date, a number of approaches have been proposed to identify differentially expressed gene sets based on microarray gene expression data. However, these methods are primarily developed for microarray data. For RNA-Seq data, the corresponding read counts are integer numbers. Since there are plenty of GSA approaches developed for microarrays, it would be expected that they are suitable for RNA-Seq data. To deal with this problem, our method converts RNA-seq data in a way of computational processing that makes the resulting transcript profiling more similar to microarray gene expression measures

We conduct simulation studies to compare our proposed method with four commonly used methods, including the Wald-Wolfowitz (WW) test (Rahmatallah et al., 2012), the Kolmogorov-Smirnov (KS) test (Rahmatallah et al., 2012), the E-Statistic (Energy) test (Székely et al., 2004), and the Quantitative Set Analysis of Gene Expression (QuSAGE) (Yaari et al., 2013). Also, we take four normalization methods into account to examine the performance of multivariate tests, including the Reads per kilobase per million (RPKM) (Mortazavi et al., 2008), the Upper-quantile normalization (UQ) (Bullard et al., 2010), the Trimmed mean of M-values (TMM) (Robinson and Oshlack, 2010), and the Log-counts per million (LCPM) (Law et al., 2014).

Consider two different phenotypes with n_1 samples for the first and n_2 samples for the second

phenotype. Each sample has measurements of same p genes. Let X_i and Y_j be p-dimensional vectors of measurements $X_i, i = 1, \ldots, n_1$ and $Y_j, j = 1, \ldots, n_2$. Suppose that X_i and Y_j are independent and identically distributed with the distribution functions F, G, mean vectors μ_x, μ_y and $p \times p$ covariance matrices Σ_x, Σ_y, respectively. We consider the problem of testing the hypothesis $H_0 : \mu_x = \mu_y$ against an alternative $H_1 : \mu_x \neq \mu_y$.

2.1 DECO: An algorithm to remove correlation in data

As we mentioned before, GSA gives an insight into expression data. It evaluates gene sets instead of individual genes. Also, it provides high statistical power and can reveal biological processes, which relates to specific phenotypes. Nevertheless, it relies on the assumption that the members of each gene set are sampled independently, which means that the genes within a gene set don't share a common biological function. Based on this assumption, it may increase false predictions.

To solve this problem, Nam (2010) proposed DECO algorithm to remove the correlation bias in the expression of each gene set in GSA. We conduct DECO and adopt the sum of square t-statistic for our later analyses. The DECO algorithm is based on the eigenvalue-decomposition of the covariance matrix of each gene set and a series of linear transformations of data.

The de-correlation procedure is described as follows.

Let X represent $p \times N$ expression data of a gene set. Here p denotes the number of genes while N denotes the total number of samples.

1. Normalize the profiles of each gene by taking log and Z-transformation, and let Y represent the transformed data.
2. Estimate the covariance matrix C of p genes from Y.
3. Apply the eigenvalue-decomposition to the positive-definite symmetric matrix C as follows: $C = UDU^{-1}$, where U is the $p \times p$ eigenvector matrix and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix with positive eigenvalues $\lambda_1, \ldots, \lambda_p$.
4. Apply the linear transformations to Y,

$$Y^* = U\sqrt{D^{-1}}U^{-1}Y,$$

where $\sqrt{D^{-1}} = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_p^{-1/2})$. This makes Y^* an identity covariance matrix.

The last step actually de-correlates the data by providing them a near spherical shape. For eigenvalues, Nam (2010) provided another two methods for them. It is possible that large gene sets have extremely small eigenvalues. If we amplify these small eigenvalues, they may lead to unstable predictions. In addition, for the coordinated patterns of gene expression, large eigenvalues are more explainable. Accordingly, the first method is based on reducing large eigenvalues by using the following truncated diagonal matrix instead of $\sqrt{D^{-1}}$,

$$\sqrt{\bar{D}^{-1}} = \mathrm{diag}(\beta_1, \ldots, \beta_p), \quad \beta_i = \begin{cases} \lambda_i^{-1/2}, & \lambda_i > T, \\ T^{-1/2}, & \text{otherwise}, \end{cases} \qquad 0 < T \leq 1.$$

Here, $T = 1$ is assumed in our tests. It implies that small eigenvalues will be rounded up to 1. To estimate the covariance matrices of gene sets accurately, the number of samples should be sufficiently large. However, in real expression data, the number of samples may not be large enough. The second method is based on a square-root de-correlation by taking one more square-root on $\sqrt{D^{-1}}$,

$$\sqrt{\tilde{D}^{-1}} = \sqrt[\gamma]{D^{-1}}, \quad \gamma = 4.$$

They denote the methods that use the truncated matrix ($\sqrt{\bar{D}^{-1}}$) as DECO-t, and the square root matrix ($\sqrt{\tilde{D}^{-1}}$) as DECO-sqrt. Moreover, we can obtain the desired data Y^* and remove the estimation error involved in the eigenvectors (U) by returning the data to the original axes. The DECO algorithm is based on the eigenvector-eigenvalue decomposition of the covariance matrix between genes in each gene set. Principal components analysis (PCA) and singular value decomposition (SVD) are two most popular dimensionality reduction techniques for gene expression data analysis to date. It allows us to capture and visually analyze the global expression patterns of gene expression data. Here the aim is to remove gene correlations, which may produce a substantial amount of variance inflation in the GSA statistics, by rescaling the principal components from the eigenvalue-decomposition of each gene set. This algorithm provides a decorrelation procedure to remove the complex gene set correlation structure and to improve the resolution of gene set analysis.

2.2 Estimating a covariance matrix with a small sample size

In practice, the number of samples may not be sufficiently large compared to the dimension of the estimated covariance matrix. Therefore, the sample covariance $S = \frac{n}{n-1} S_{\mathrm{ML}}$ and the standard maximum likelihood estimator S_{ML} cannot provide a good approximation of the true covariance matrix. Moreover, when estimating large gene sets, S and S_{ML} may produce eigenvalues that equal to zero. It implies that they can't always provide positive-definite eigenvalues. In the Step 2 of the DECO algorithm, the shrinkage covariance estimator (Schäfer and Strimmer, 2005) can be used to estimate the covariance matrices instead of the sample covariance because it provides a more accurate estimate which is always positive-definite. With the shrinkage covariance estimator, all of the eigenvalues become positive.

Shrinkage covariance estimation is well-conditioned and always positive-definite even with small number of samples. The purpose of it is variance reduction. Take a look at the bias-variance decomposition of the mean squared error (MSE) for the sample covariance S,

$$\mathrm{MSE}(S) = \mathrm{Bias}(S)^2 + \mathrm{Var}(S).$$

When $\mathrm{Bias}(S) = 0$, the only way to decrease the MSE of S is reducing its variance. The shrinkage estimator $S^* = [s^*_{ij}]$ is defined as follows,

$$s^*_{ij} = \begin{cases} s_{ii}, & i = j, \\ r^*_{ij}\sqrt{s_{ii}s_{jj}}, & \text{otherwise}, \end{cases}$$

$$r^*_{ij} = \begin{cases} 1, & i = j, \\ r_{ij} \max\left\{0, \min(1, \bar{\lambda}^*)\right\}, & \text{otherwise}, \end{cases}$$

$$\text{with} \quad \bar{\lambda}^* = \frac{\sum_{i \neq j} \bar{\mathrm{Var}}(r_{ij})}{\sum_{i \neq j} r^2_{ij}},$$

where s_{ii} and r_{ij} denote the empirical variance and correlation respectively. Assume that x_{ki} is the i-th gene in the standardized k-th observation. Define $w_{kij} = x_{ki}x_{kj}$ and $\bar{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$. This gives $\bar{\mathrm{Var}}(r_{ij}) = \frac{n}{(n-1)^3}(w_{kij} - \bar{w}_{ij})^2$.

2.3 Gene Set Analysis with DECO

Suppose there are p genes and N samples in each gene set. As we mentioned before, self-contained approaches are all based on random permutation of sample labels to assess the significance of

each test statistic. Here, we demonstrate how GSA works with DECO algorithm.

1. DECO the gene set.
2. Compute the Welch's t-statistic of each individual gene, t_i.
3. Combine the individual t-statistic within the gene set into a summarized statistic using the sum of square t-statistic: $T_{obs} = \sum_{i=1}^{p} t_i^2/p$.
4. Randomly permute the sample labels of the transformed data Y^* B times.
5. Compute permuted sum of square t-statistics, $T^*_{(b)}$; $b = 1, \ldots, B$.
6. Calculate the empirical p-value, $\mathrm{P} - \mathrm{value} = \sum_{b=1}^{B} I\left[T^*_{(b)} \geq T_{obs}\right]/B$.

With the observed sum of square t-statistic and the permuted sum of square t-statistics, we can calculate the p-value of each gene set. It is important to note that by applying the de-correlation procedure to gene expression data does not transform the data into independent samples except that the gene expression profiles follow a multivariate Gaussian distribution. Fortunately, gene log2 expression levels are highly correlated and, very likely, have approximately normal distribution (Zhang et al., 2018). Therefore, it seems reasonable to use the de-correlation procedure for such data to avoid overfitting when using sample permutations GSA methods.

3 Results

3.1 Simulation Study

For count data, we considered two scenarios. One is that the genes of each gene set are sampled independently; the other is that some of the genes in each gene set are correlated. The former is called independent data and the latter is called correlated data. We generated count data from Negative Binomial (NB) distribution for RNA-Seq experiments with mean counts μ and dispersion parameter ϕ. Let Y_{ij} denote the count for gene i in sample j, then $Y_{ij} \sim \mathrm{NB}\{\mathrm{mean} = \mu_{ij}, \mathrm{var} = \mu_{ij}(1 + \mu_{ij}\phi_{ij})\} = \mathrm{NB}(\mu_{ij}, \phi_{ij})$. In addition, for each gene, we used the Bioconductor package edgeR to estimate the dispersion parameter.

Prior to GSA analyses, we used different normalization methods to normalize read counts. Then, we transformed them to log-scale by using the transformation function $\log_2(1 + Y_{ij})$. In order to evaluate the performance of different multivariate tests accurately, two parameters, γ and FC, were considered in our analyses. The γ parameter is the percentage of genes truly differentially expressed in a gene set; the FC parameter denotes the amount of fold change in read counts between two phenotypes. We calculated mean counts, μ_i, and estimated dispersion parameter, ϕ_i, from Pickrell dataset for each gene i. We describe our simulation study design by the following steps.

1. Randomly select numbers of genes from Pickrell dataset without replacement. The pairs of μ_i and ϕ_i will be true parameters for simulated gene i.
2. Randomly select simulated gene i to be either differentially expressed (DE) gene or non-expressed gene between two phenotypes. Furthermore, we pick half of the DE genes to be up-regulated and half to be down-regulated between two phenotypes to avoid having all of the DE genes are up-regulated for all generated gene sets.
3. Set a log fold change between two phenotypes.
4. Construct a correlation structure into the differential expression pattern of simulated gene i.
5. Generate the count of each simulated gene i in sample j from $\mathrm{NB}(\mu_{ij}, \phi_i)$.
6. If the counts of simulated gene i are all zero for all samples, then we return to step 1 to regenerate the count.

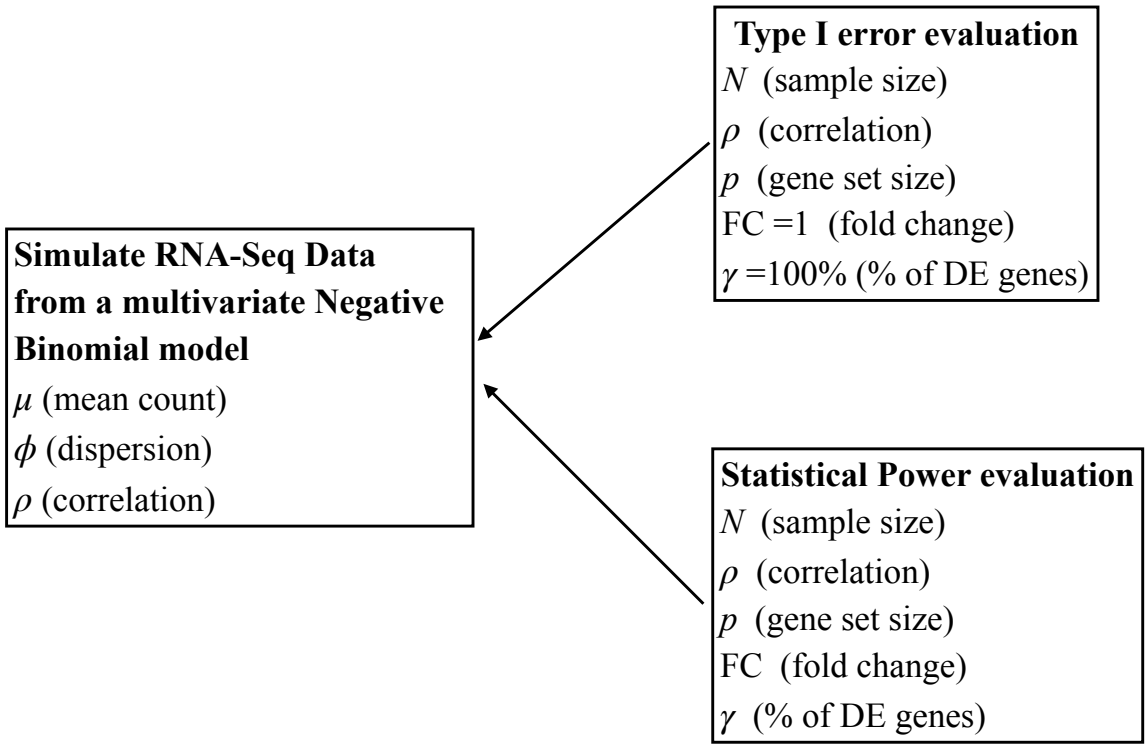

Figure 1: Simulation scheme used to generate RNA-Seq data with a correlation structure and known parameters.

In steps 4 and 5 above, we use the following algorithm to generate multivariate Negative Binomial random variables.

1. Generate a p-dimensional Normal vector $\vec{X}$ with mean vector $\vec{\mu} = \mathbf{0}$, variance vector $\vec{\sigma} = \mathbf{1}$ and a correlation matrix.
2. For each element $X_i, i = 1, 2, \ldots, p$, calculate the Normal cumulative distribution function (CDF) : $\Phi(X_i)$.
3. For each $\Phi(X_i)$, calculate the Negative Binomial inverse CDF with dispersion ϕ_i and mean μ_i. Let Ξ be Negative Binomial CDF, then $X_i^{\mathrm{NB}} = \Xi^{-1}(\Phi(X_i))$.

For our correlation matrix, we set an equal correlation between any two genes. By setting different simulated parameters, we can get different pseudo-datasets. For all of simulations, we assumed a significance level of $\alpha = 0.05$. Figure 1 outlines the parameters used to simulate the correlated RNA-Seq data for performance evaluation.

To estimate the Type I error rates for all multivariate tests using simulated counts data, we set γ and FC to 1 and generated two datasets of equal sample size, $N/2, N \in \{20, 40, 60\}$ with 3000 genes, 20 times. For each time, we randomly selected 50 gene sets with numbers of genes $p, p \in \{16, 60, 100\}$, forming 1000 gene sets. Then we estimated the proportion of gene sets that rejected $H_0 : \mu_x = \mu_y$ among the 1000 gene sets. We estimated the Type I error rates of DECO algorithm combined with E-stat and called this method as DECO-E. We denoted the method combining DECO-t with E-stat as DECOt-E, and the method combining DECO-sqrt with E-stat as DECOs-E. For convenience, DECO, DECO-t and DECO-sqrt refer to our proposed methods combining the sum of square t-statistic with DECO, DECO-t and DECO-sqrt respectively in

Table 1: Type I error rates of all methods for independent data.

		$p = 16$				$p = 60$				$p = 100$			
		RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM
$N = 20$	E-stat	0.044	0.051	0.047	0.054	0.049	0.053	0.049	0.062	0.042	0.050	0.053	0.045
	WW	0.081	0.094	0.072	0.079	0.088	0.081	0.079	0.102	0.099	0.097	0.086	0.079
	KS	0.087	0.090	0.086	0.088	0.084	0.089	0.075	0.096	0.086	0.088	0.089	0.069
	QuSAGE	0.048	0.051	0.048	0.053	0.055	0.055	0.057	0.054	0.041	0.054	0.051	0.036
	DECO	0.052	0.050	0.048	0.050	0.048	0.046	0.048	0.050	0.048	0.048	0.050	0.055
	DECO-sqrt	0.049	0.047	0.048	0.051	0.048	0.046	0.048	0.048	0.046	0.049	0.050	0.055
	DECO-t	0.051	0.051	0.050	0.052	0.048	0.044	0.046	0.052	0.050	0.054	0.053	0.052
	DECO-E	0.054	0.047	0.048	0.053	0.045	0.052	0.045	0.048	0.046	0.048	0.051	0.056
	DECOs-E	0.046	0.050	0.052	0.051	0.050	0.052	0.049	0.062	0.046	0.048	0.051	0.048
	DECOt-E	0.049	0.046	0.052	0.055	0.047	0.049	0.044	0.064	0.043	0.041	0.048	0.042
$N = 40$	E-stat	0.051	0.046	0.049	0.052	0.043	0.042	0.042	0.047	0.058	0.054	0.050	0.046
	WW	0.078	0.069	0.092	0.080	0.072	0.079	0.074	0.068	0.078	0.073	0.071	0.081
	KS	0.080	0.071	0.064	0.064	0.072	0.075	0.057	0.069	0.067	0.073	0.076	0.068
	QuSAGE	0.050	0.050	0.049	0.050	0.051	0.057	0.055	0.038	0.047	0.041	0.042	0.042
	DECO	0.056	0.051	0.051	0.058	0.057	0.062	0.055	0.053	0.047	0.048	0.050	0.045
	DECO-sqrt	0.056	0.050	0.051	0.057	0.057	0.060	0.056	0.053	0.045	0.048	0.050	0.045
	DECO-t	0.053	0.050	0.053	0.058	0.057	0.060	0.056	0.054	0.043	0.049	0.048	0.043
	DECO-E	0.054	0.052	0.051	0.058	0.061	0.062	0.058	0.059	0.050	0.050	0.049	0.040
	DECOs-E	0.052	0.055	0.058	0.051	0.047	0.046	0.049	0.045	0.055	0.049	0.056	0.037
	DECOt-E	0.053	0.049	0.050	0.050	0.049	0.045	0.046	0.047	0.054	0.053	0.047	0.050
$N = 60$	E-stat	0.049	0.052	0.056	0.048	0.052	0.051	0.048	0.052	0.037	0.030	0.037	0.043
	WW	0.063	0.075	0.078	0.068	0.069	0.059	0.066	0.060	0.069	0.075	0.060	0.063
	KS	0.059	0.062	0.062	0.059	0.057	0.060	0.059	0.050	0.054	0.058	0.048	0.067
	QuSAGE	0.038	0.040	0.034	0.055	0.048	0.048	0.045	0.057	0.065	0.065	0.063	0.054
	DECO	0.051	0.049	0.049	0.052	0.046	0.041	0.043	0.044	0.042	0.047	0.045	0.050
	DECO-sqrt	0.052	0.049	0.048	0.052	0.045	0.043	0.042	0.044	0.043	0.047	0.044	0.050
	DECO-t	0.050	0.050	0.050	0.051	0.043	0.044	0.043	0.047	0.043	0.048	0.044	0.049
	DECO-E	0.049	0.050	0.049	0.051	0.044	0.042	0.042	0.046	0.039	0.036	0.043	0.045
	DECOs-E	0.058	0.062	0.061	0.051	0.053	0.056	0.048	0.051	0.032	0.031	0.033	0.038
	DECOt-E	0.050	0.053	0.056	0.053	0.053	0.053	0.051	0.054	0.034	0.035	0.037	0.038

later analyses.

Table 1 presents the simulation results of independent data for the Type I error rates. We found that all multivariate tests provide good estimates except for WW and KS. When the sample size (N) increases, the Type I error rates of WW and KS decrease obviously. In addition, compared with other tests, WW seems to be the most liberal and the KS is in the second place. Also, TMM almost provides more conservative estimates than other normalization methods and LCPM tends to have the most conservative or the most liberal estimates among these four normalization methods. For correlated data, we added some correlation information to our simulated counts data. Among 3000 genes, there are 600 genes with correlation (cor $= 0.3, 0.6, 0.9$). Tables 2-4 show the Type I error rates of all tests with normalizations under different correlation strengths. In Table 2 cor $= 0.3$, the result is quite similar to the case of independent data. In Table 3 cor $= 0.6$, we found that all tests control the Type I error rate well under different gene set size when $N = 60$. As for Table 4 cor $= 0.9$, the most conservative estimates occurred at the conditions when $N = 40$ and $p = 60$ while the most liberal estimates

Table 2: Type I error rates of all methods for correlated data (cor=0.3).

cor = 0.3		$p = 16$				$p = 60$				$p = 100$			
		RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM
$N = 20$	E-stat	0.050	0.049	0.052	0.050	0.041	0.043	0.041	0.043	0.034	0.039	0.036	0.050
	WW	0.077	0.084	0.082	0.083	0.082	0.069	0.074	0.067	0.069	0.066	0.075	0.097
	KS	0.086	0.088	0.093	0.103	0.078	0.080	0.089	0.077	0.086	0.083	0.093	0.079
	QuSAGE	0.051	0.051	0.050	0.060	0.048	0.054	0.054	0.053	0.035	0.051	0.043	0.028
	DECO	0.043	0.044	0.047	0.045	0.048	0.047	0.051	0.051	0.040	0.043	0.044	0.043
	DECO-sqrt	0.043	0.043	0.046	0.046	0.046	0.047	0.052	0.050	0.035	0.038	0.041	0.043
	DECO-t	0.044	0.045	0.050	0.047	0.047	0.044	0.051	0.049	0.043	0.038	0.040	0.040
	DECO-E	0.047	0.044	0.051	0.050	0.037	0.041	0.041	0.042	0.036	0.039	0.036	0.042
	DECOs-E	0.053	0.050	0.048	0.048	0.041	0.043	0.040	0.033	0.038	0.038	0.036	0.048
	DECOt-E	0.051	0.052	0.048	0.050	0.038	0.043	0.043	0.040	0.041	0.042	0.040	0.050
$N = 40$	E-stat	0.052	0.054	0.049	0.053	0.045	0.047	0.043	0.055	0.058	0.055	0.052	0.046
	WW	0.057	0.064	0.058	0.076	0.075	0.078	0.073	0.066	0.070	0.067	0.066	0.068
	KS	0.059	0.071	0.073	0.049	0.070	0.066	0.066	0.077	0.064	0.062	0.066	0.063
	QuSAGE	0.059	0.058	0.059	0.051	0.069	0.059	0.062	0.057	0.051	0.046	0.060	0.055
	DECO	0.051	0.048	0.046	0.055	0.051	0.047	0.044	0.046	0.063	0.058	0.061	0.068
	DECO-sqrt	0.051	0.048	0.048	0.055	0.046	0.045	0.047	0.046	0.061	0.057	0.063	0.065
	DECO-t	0.052	0.047	0.049	0.056	0.045	0.046	0.044	0.045	0.061	0.054	0.061	0.062
	DECO-E	0.051	0.053	0.050	0.062	0.044	0.046	0.044	0.038	0.061	0.062	0.060	0.055
	DECOs-E	0.048	0.050	0.045	0.056	0.046	0.048	0.040	0.052	0.064	0.064	0.062	0.048
	DECOt-E	0.057	0.055	0.048	0.056	0.047	0.046	0.048	0.051	0.053	0.051	0.053	0.050
$N = 60$	E-stat	0.047	0.049	0.048	0.053	0.040	0.036	0.044	0.041	0.053	0.055	0.051	0.044
	WW	0.073	0.065	0.061	0.065	0.066	0.052	0.060	0.069	0.069	0.071	0.055	0.063
	KS	0.048	0.057	0.052	0.065	0.051	0.062	0.069	0.055	0.061	0.073	0.059	0.047
	QuSAGE	0.058	0.049	0.045	0.052	0.045	0.051	0.041	0.044	0.043	0.045	0.059	0.054
	DECO	0.056	0.058	0.062	0.047	0.038	0.040	0.040	0.041	0.059	0.059	0.060	0.066
	DECO-sqrt	0.055	0.057	0.059	0.047	0.037	0.039	0.039	0.042	0.059	0.061	0.061	0.074
	DECO-t	0.052	0.055	0.056	0.047	0.038	0.040	0.040	0.042	0.061	0.062	0.057	0.077
	DECO-E	0.055	0.058	0.050	0.053	0.035	0.037	0.035	0.038	0.055	0.053	0.057	0.058
	DECOs-E	0.054	0.056	0.053	0.052	0.039	0.040	0.038	0.038	0.045	0.046	0.048	0.047
	DECOt-E	0.045	0.048	0.049	0.061	0.041	0.040	0.043	0.040	0.054	0.056	0.049	0.044

occurred at the conditions when $N = 20$ and $p = 100$. As expected, the Type I error rates decrease as the sample size increases. Also, the Type I error rate is unrelated to the increasing number of genes. Intuitively, when the correlation between genes increases, it seems that all test methods are unstable to correctly estimate the Type I error rates.

To estimate the empirical power for all multivariate tests using simulated count data, we considered $\gamma \in \left(\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\right)$ and FC from 1.2 to 2.8. Also, we set true expression rate as 0.1. We estimated the empirical power by testing the null hypothesis $H_0 : FC = 1$ against the alternative hypothesis $H_1 : FC \neq 1$ when H_1 holds true. For independent data, among 3000 genes, 300 genes were differentially expressed and the rest 2700 genes were non-differentially expressed. Figure 2 illustrates the empirical power of all tests with RPKM normalization when $p = 16$, $N = 20, 40, 60$ and $\gamma = 0.125, 0.25, 0.5$. The similar results for UQ, TMM, LCPM normalizations can be found in the Additional file (Figures S1-3). It implies that the power increases as N and γ increase. Moreover, DECO, DECO-sqrt, and DECO-t outperform other multivariate tests, followed by DECO-E and DECOs-E. In addition, TMM normalization almost leads to the highest power

Table 3: Type I error rates of all methods for correlated data (cor=0.6).

cor = 0.6		$p = 16$				$p = 60$				$p = 100$			
		RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM
$N = 20$	E-stat	0.043	0.037	0.048	0.037	0.070	0.071	0.064	0.072	0.059	0.063	0.057	0.062
	WW	0.082	0.082	0.087	0.080	0.091	0.090	0.093	0.091	0.118	0.099	0.109	0.076
	KS	0.102	0.108	0.111	0.085	0.095	0.090	0.133	0.101	0.108	0.127	0.091	0.096
	QuSAGE	0.046	0.047	0.046	0.050	0.047	0.050	0.050	0.053	0.061	0.076	0.061	0.047
	DECO	0.043	0.043	0.043	0.042	0.083	0.074	0.073	0.082	0.079	0.078	0.071	0.071
	DECO-sqrt	0.049	0.044	0.045	0.042	0.079	0.077	0.071	0.081	0.077	0.077	0.074	0.079
	DECO-t	0.048	0.045	0.045	0.041	0.076	0.081	0.075	0.079	0.074	0.073	0.073	0.074
	DECO-E	0.045	0.048	0.048	0.049	0.075	0.072	0.076	0.071	0.068	0.066	0.067	0.075
	DECOs-E	0.042	0.036	0.045	0.041	0.070	0.073	0.074	0.067	0.070	0.067	0.065	0.061
	DECOt-E	0.045	0.047	0.050	0.035	0.066	0.066	0.066	0.075	0.056	0.058	0.055	0.061
$N = 40$	E-stat	0.052	0.053	0.059	0.041	0.054	0.059	0.061	0.061	0.066	0.063	0.071	0.067
	WW	0.076	0.075	0.065	0.069	0.065	0.070	0.067	0.087	0.064	0.070	0.070	0.061
	KS	0.066	0.065	0.055	0.062	0.064	0.073	0.080	0.067	0.085	0.067	0.069	0.056
	QuSAGE	0.060	0.049	0.054	0.048	0.039	0.043	0.046	0.041	0.076	0.076	0.057	0.066
	DECO	0.046	0.047	0.041	0.047	0.071	0.068	0.061	0.064	0.079	0.067	0.065	0.067
	DECO-sqrt	0.048	0.045	0.045	0.045	0.076	0.071	0.071	0.067	0.077	0.067	0.066	0.084
	DECO-t	0.050	0.046	0.045	0.046	0.072	0.072	0.071	0.068	0.084	0.077	0.077	0.087
	DECO-E	0.043	0.046	0.043	0.042	0.065	0.061	0.064	0.066	0.068	0.059	0.059	0.061
	DECOs-E	0.045	0.051	0.050	0.039	0.060	0.065	0.064	0.060	0.069	0.063	0.067	0.069
	DECOt-E	0.055	0.052	0.058	0.041	0.058	0.061	0.063	0.062	0.062	0.058	0.060	0.068
$N = 60$	E-stat	0.056	0.057	0.055	0.048	0.046	0.042	0.046	0.055	0.045	0.044	0.044	0.054
	WW	0.065	0.062	0.055	0.065	0.061	0.070	0.063	0.086	0.065	0.069	0.054	0.064
	KS	0.063	0.056	0.066	0.066	0.052	0.056	0.063	0.046	0.061	0.060	0.059	0.058
	QuSAGE	0.055	0.053	0.055	0.058	0.052	0.052	0.055	0.051	0.041	0.039	0.048	0.039
	DECO	0.043	0.045	0.041	0.041	0.038	0.040	0.034	0.045	0.033	0.033	0.033	0.047
	DECO-sqrt	0.044	0.045	0.042	0.044	0.036	0.037	0.034	0.041	0.036	0.032	0.037	0.036
	DECO-t	0.047	0.041	0.043	0.052	0.039	0.038	0.037	0.038	0.049	0.041	0.039	0.035
	DECO-E	0.044	0.046	0.044	0.047	0.046	0.044	0.045	0.044	0.038	0.037	0.037	0.043
	DECOs-E	0.053	0.053	0.053	0.040	0.043	0.046	0.043	0.050	0.046	0.043	0.044	0.052
	DECOt-E	0.061	0.057	0.060	0.044	0.047	0.046	0.043	0.047	0.045	0.047	0.047	0.054

while LCPM gives the lowest power performance among four normalization methods. When the gene set size increases to $p = 60$ (Figure 3) and for $p = 100$ (Figure 4), we compared power performance for all tests with TMM normalization. Both scenarios show the similar performance. Overall, our approach DECO-t has the highest power performance of all the methods across all scenarios. In addition, when p and γ increase, the power increases in detecting small fold changes for all tests.

For correlated data, among 3000 genes, 300 genes were differentially expressed with correlation, 300 genes were non-differentially expressed with correlation, and the rest 2400 genes were non-differentially expressed without correlation (cor $= 0.3, 0.6, 0.9$). Figure 5 shows differences in power performance across all methods using TMM normalization, $p = 60$ and cor $= 0.6$. It is clear that as N and γ increase, the power increases. All simulation results can be found in the Additional file (Figure S4-23) by varying normalization method, gene set size and correlation level among genes. Overall, our proposed methods DECO and DECO-sqrt show the best results in the simulations, with higher power performance in all scenarios. QuSAGE could lead

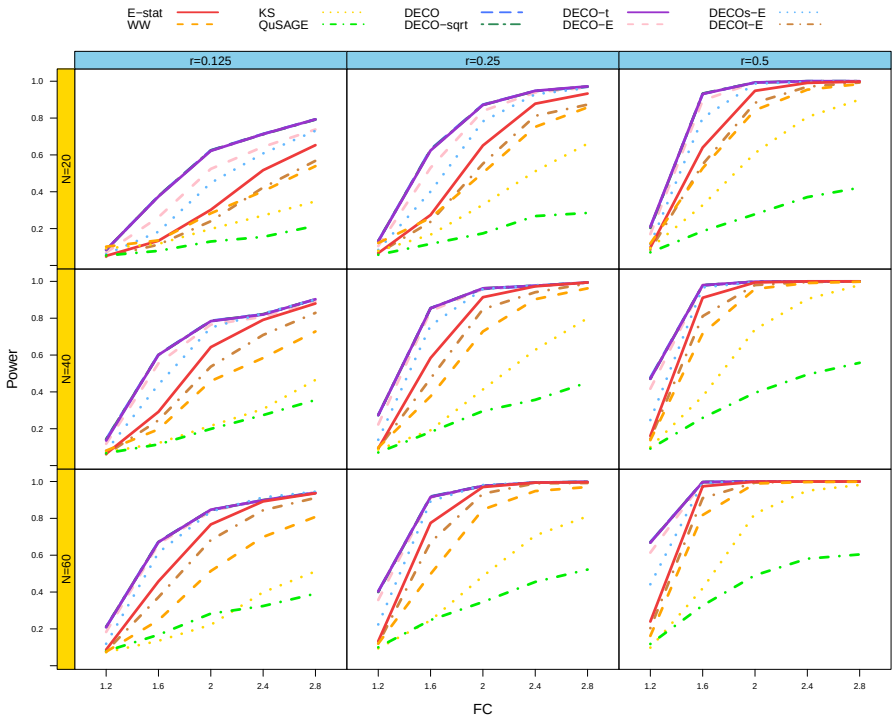

Figure 2: The power curves of multivariate tests for independent data with gene size ($p = 16$) and RPKM normalization.

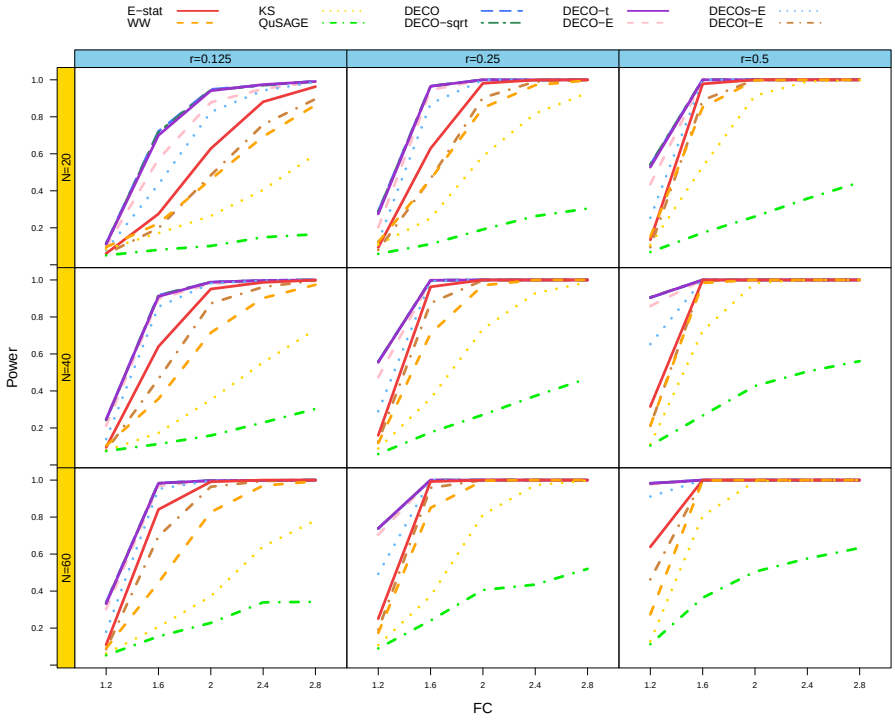

Figure 3: The power curves of multivariate tests for independent data with gene size ($p = 60$) and TMM normalization.

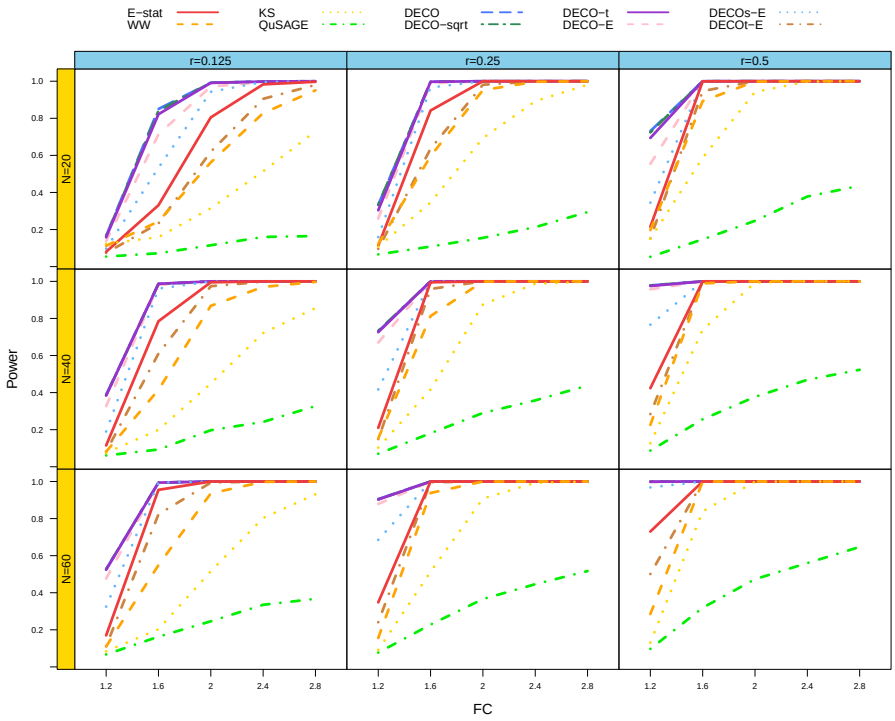

Figure 4: The power curves of multivariate tests for independent data with gene size ($p = 100$) and TMM normalization.

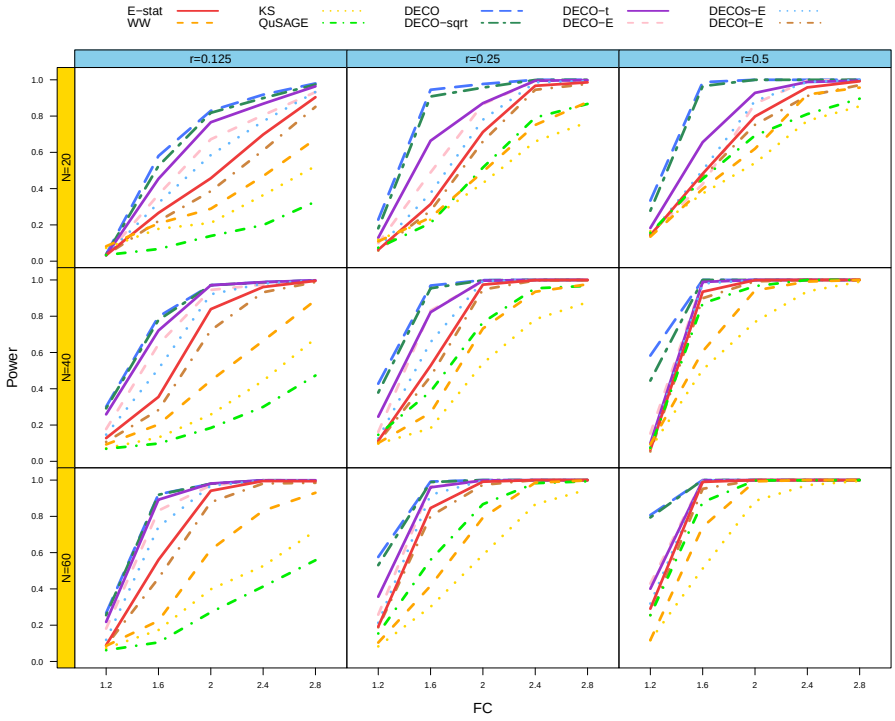

Figure 5: The power curves of multivariate tests for correlated data ($cor = 0.6$) with gene size ($p = 60$) and TMM normalization.

Table 4: Type I error rates of all methods for correlated data (cor=0.9).

cor = 0.9		$p = 16$				$p = 60$				$p = 100$			
		RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM	RPKM	UQ	TMM	LCPM
$N = 20$	E-stat	0.047	0.046	0.043	0.052	0.033	0.037	0.034	0.026	0.078	0.075	0.075	0.052
	WW	0.074	0.083	0.074	0.075	0.079	0.070	0.077	0.072	0.090	0.085	0.089	0.093
	KS	0.086	0.096	0.090	0.091	0.069	0.075	0.069	0.085	0.101	0.088	0.083	0.104
	QuSAGE	0.043	0.052	0.045	0.049	0.041	0.044	0.042	0.054	0.064	0.060	0.056	0.050
	DECO	0.038	0.039	0.038	0.039	0.030	0.034	0.026	0.032	0.065	0.066	0.070	0.066
	DECO-sqrt	0.038	0.035	0.037	0.037	0.031	0.031	0.025	0.024	0.082	0.085	0.087	0.063
	DECO-t	0.041	0.038	0.040	0.036	0.021	0.024	0.022	0.021	0.100	0.102	0.100	0.091
	DECO-E	0.038	0.040	0.036	0.039	0.029	0.029	0.031	0.021	0.069	0.068	0.066	0.057
	DECOs-E	0.040	0.038	0.042	0.048	0.021	0.023	0.021	0.022	0.078	0.076	0.071	0.055
	DECOt-E	0.044	0.046	0.043	0.053	0.035	0.033	0.033	0.025	0.074	0.077	0.073	0.051
$N = 40$	E-stat	0.051	0.052	0.050	0.059	0.029	0.030	0.027	0.033	0.051	0.051	0.054	0.066
	WW	0.067	0.061	0.064	0.066	0.061	0.067	0.063	0.057	0.080	0.081	0.080	0.085
	KS	0.055	0.073	0.056	0.065	0.045	0.059	0.066	0.045	0.064	0.080	0.073	0.081
	QuSAGE	0.038	0.046	0.045	0.048	0.047	0.039	0.039	0.051	0.060	0.067	0.057	0.050
	DECO	0.068	0.065	0.062	0.052	0.045	0.041	0.043	0.042	0.048	0.055	0.057	0.047
	DECO-sqrt	0.068	0.066	0.062	0.060	0.037	0.041	0.039	0.050	0.055	0.059	0.059	0.048
	DECO-t	0.067	0.063	0.061	0.063	0.032	0.033	0.032	0.040	0.062	0.066	0.064	0.062
	DECO-E	0.058	0.055	0.056	0.061	0.040	0.044	0.042	0.036	0.059	0.061	0.059	0.057
	DECOs-E	0.063	0.063	0.062	0.051	0.028	0.027	0.024	0.034	0.051	0.053	0.051	0.062
	DECOt-E	0.051	0.051	0.051	0.056	0.034	0.028	0.030	0.037	0.056	0.054	0.057	0.065
$N = 60$	E-stat	0.053	0.058	0.054	0.061	0.086	0.084	0.086	0.071	0.051	0.050	0.056	0.038
	WW	0.072	0.074	0.066	0.072	0.059	0.055	0.058	0.061	0.074	0.075	0.076	0.068
	KS	0.075	0.063	0.051	0.058	0.063	0.066	0.064	0.073	0.055	0.052	0.047	0.054
	QuSAGE	0.067	0.070	0.061	0.051	0.041	0.052	0.040	0.046	0.035	0.035	0.031	0.042
	DECO	0.053	0.054	0.058	0.055	0.077	0.078	0.076	0.074	0.045	0.049	0.051	0.048
	DECO-sqrt	0.053	0.054	0.059	0.052	0.074	0.082	0.083	0.074	0.052	0.052	0.054	0.053
	DECO-t	0.052	0.052	0.050	0.050	0.085	0.083	0.082	0.078	0.059	0.059	0.061	0.060
	DECO-E	0.052	0.051	0.055	0.045	0.078	0.077	0.073	0.072	0.047	0.044	0.044	0.042
	DECOs-E	0.063	0.058	0.061	0.059	0.081	0.085	0.081	0.076	0.049	0.052	0.054	0.040
	DECOt-E	0.056	0.059	0.053	0.059	0.079	0.079	0.078	0.080	0.056	0.055	0.056	0.039

to the lowest power of test when p and γ are small. Interestingly, when the correlation increases, both our proposed methods DECO and DECO-sqrt still perform well, but the powers of all tests decrease slightly. When p increases, the power increases for all tests even with small fold changes. In addition, it appears that there are no significant differences in power performance across RPKM, UQ, and TMM normalizations. However, the LCPM normalization appears to be relatively less powerful in all simulations.

3.2 Real data analysis

In this section, we present the performance of the GSA methods on a real dataset using four normalization methods. We re-analyzed the Pickrell dataset from the Bioconductor package tweeDEseqCountData. The dataset consists of 69 lymphoblastoid cell lines (LCL), which were derived from Yoruban Nigerian individuals. The annotEnsembl63 is a data frame object with annotation data for the human genes. Our gene information was obtained from the annotEnsembl63

in tweeDEseqCountData package. Then, we discarded genes that didn't have gene length information. Also, any genes with an average count per million (cpm) less than 0.1 were discarded. Gene sets were obtained from the C2 curated pathways of the molecular signatures database (MSigDB) 3.0. The GSVAdata package provides the list of these gene sets. Next, we used the org.Hs.eg.db package to convert the entrez identifiers to gene symbol identifiers. In addition, genes which did not exist in the Pickrell dataset were discarded from the C2 curated pathways. Last, we kept gene sets with the number of genes between 10 and 500. We randomly chose 29 samples for each gender. The dataset was left with 4741 genes and 58 samples, while the resulted pathways were left with 1085 pathways.

We detect pathways whether they are differentially expressed (DE) between male and female samples using multivariate tests (E-stat, WW, KS, QuSAGE, and DECO-sqrt) by using different normalization methods. Also, we set a significance level $\alpha = 0.05$. To summarize, DECO-sqrt is the most liberal test while QuSAGE is the most conservative test. Among four normalization methods, DECO-sqrt detects 196 DE pathways with RPKM, 174 DE pathways with UQ, 158 DE pathways with TMM, and 221 DE pathways with LCPM; QuSAGE detects a total of 346 pathways. Figure 6 presents the Venn diagrams for the TMM normalization. The performance of these methods is very similar for other normalization methods (Additional file, Figures S24-26). For normalization methods, TMM and LCPM detect more pathways compared to others. Figure 7 shows the common pathways detected by our proposed method DECO-sqrt for different normalization methods. We found that DECO-sqrt detects much more common pathways than others across four normalizations. Corresponding results for other competing methods can be found in the Additional file, Figures S27-30.

To have more information about DECO-sqrt's ability in detecting differentially expressed gene sets, we calculate the number of DE gene sets detected only by DECO-sqrt. We found that there are 56 DE gene sets detected only by DECO-sqrt. In order to know whether there is any important gene set among them, we compared these 56 gene sets with the results presented by Zhang et al. (2009). They used GSEA and C2 curated pathways to reveal differences in YRI (Yoruba people from Ibadan, Nigeria) expression data between males and females. The YRI expression data was derived from the Affymetrix GeneChip Human Exon 1.0 ST array (exon array) dataset. We found one gene set, MANALO_HYPOXIA_DN, in common. The gene set is enriched in female samples and related to hypoxia and overexpression.

According to Brannon et al. (2012), they identified whether gender influences on tumor biology. Their materials are C2 curated gene sets and T261 array. We found that there are 9 gene sets in common (2 DE gene sets enriched in females and 7 DE gene sets enriched in males). See Table 5.

4 Conclusions

Identifying differentially expressed genes is a useful way to overview genome wide expression profiling. In most cases, the genes within each gene set aren't distributed independently, which does not agree with the assumption of most statistical tests. Therefore, making this independence assumption may lead to difficulties of biological interpretation by conventional GSA methods. In this paper, we used the DECO algorithm to alleviate this problem by removing the correlations in each gene set. The proposed method could reduce the estimation error for the covariance matrix of each gene set because the intra-gene set correlation has a significant impact on the distribution of testing statistics.

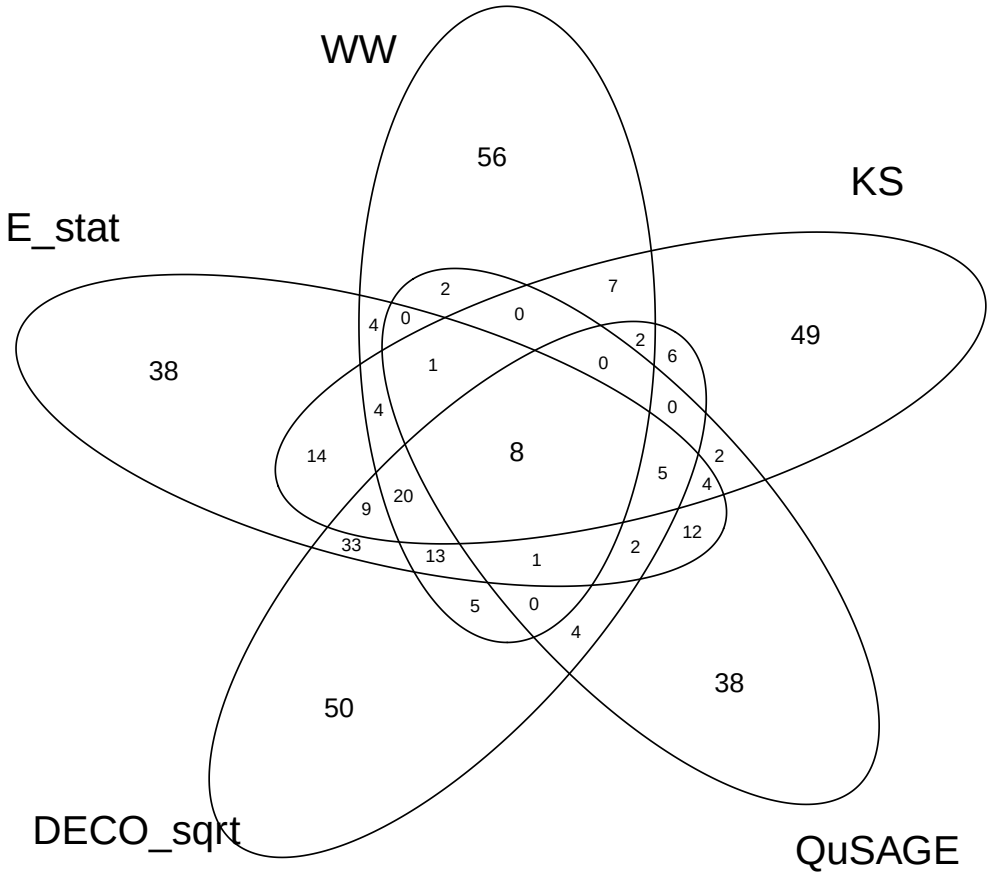

Figure 6: The Venn diagram of the C2 curated pathways detected in the Pickrell dataset by different tests under the TMM normalization.

We compare and evaluate our proposed method to other existing self-contained GSA approaches in combination with four different normalizations (RPKM, TMM, UQ, and LCPM) on simulated and real RNA-Seq data. The results show that our proposed methods appropriately control the type I error and have the highest power. These observations indicate that intra-gene set correlation should be taken into account when evaluating the gene set enrichment significance. Based on these results, we observed that all tests with TMM normalization have the highest statistical power, while tests with LCPM normalization have the smallest power and the smallest Type I error rates. When the gene set size increases to 100, the Type I error rate and the power depend only on the test statistics and are insensitive to the different normalizations. Overall, DECO and DECO-sqrt have the highest power of all the methods with intermediate to large correlation and DECO-t performs well with independent and small correlation across all scenarios. In addition, DECO and DECO-sqrt are more powerful in detecting smaller fold changes in the simulation study. From the real RNA-Seq dataset, very often different normalization and GSA methods may detect different differentially expressed gene sets. It reveals that every method is constructed under different assumptions and searches for the significant gene sets from different perspective. As expected, our proposed approach is much powerful and detects more common pathways than others across four normalizations. The experiments conducted here support the application of de-correlation algorithm combining with univariate statistics, recommending that

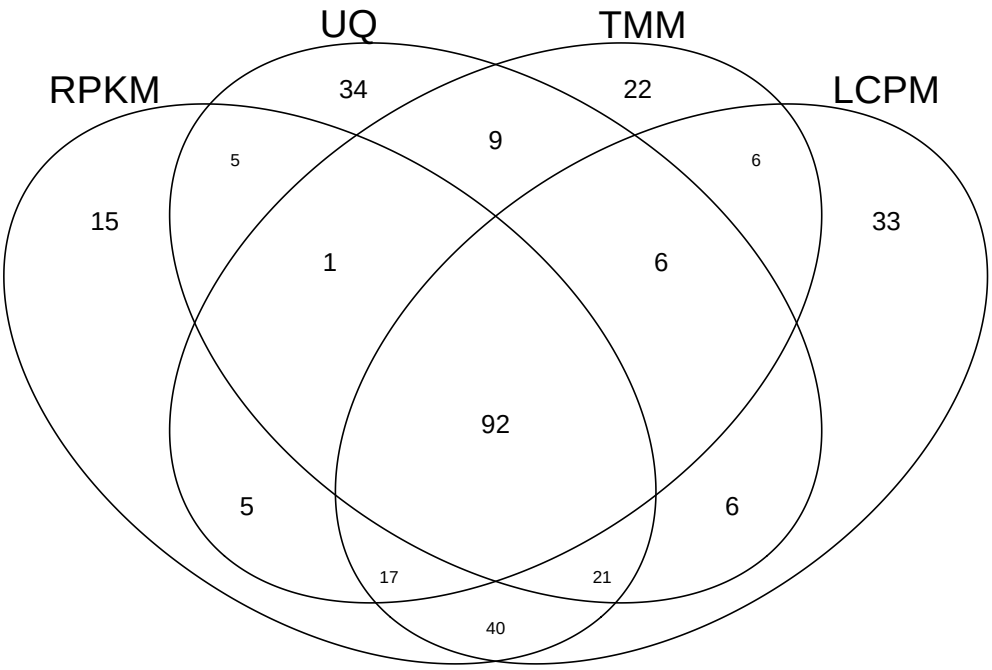

Figure 7: The Venn diagram of the C2 curated pathways detected in the Pickrell dataset by the DECO-sqrt method under varying normalization methods.

Table 5: Tumor-related gene sets.

Gene Set Name	Gender
FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN	Females
ELVIDGE_HYPOXIA_DN	Females
TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_HSC_DN	Males
ZHAN_MULTIPLE_MYELOMA_CD1_UP	Males
ODONNELL_TFRC_TARGETS_UP	Males
RODWELL_AGING_KIDNEY_NO_BLOOD_UP	Males
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	Males
TSAI_RESPONSE_TO_IONIZING_RADIATION	Males
TARTE_PLASMA_CELL_VS_B_LYMPHOCYTE_DN	Males

both DECO and DECO-sqrt methods were successful in identifying differentially expressed gene sets from RNA-Seq data.

An important conclusion from our work reveals that normalization methods are not robust to the intra-gene set correlation. This implies that differentially expressed gene sets have low reproducibility from normalization methods. In fact an increasing number of genomic data by different types of available platforms suffer from the same problem as the examples presented here, having few overlapping gene sets identified simultaneously by different normalization methods. These results suggest that reproducibility can be improved by optimizing experimental designs, randomizing potential experimental factors to biological samples, and increasing the sample size. Also, it is worth noticing that the selection difference between various GSA approaches is due to the null hypothesis which they test for significant gene sets from different perspective. To

reduce the burden of multiple testing and relatively small sample size, a potential improvement can be reached to adapt the GSA approaches to integrate analyses from multiple studies using a meta-analysis.

The proposed strategy can also be applied to other sorts of high-throughput genomic data when the intra-gene set correlation cannot be ignorable. This consideration of intra-gene set correlation provides a more significant improvement to the statistical power. Indeed, many alternative GSA approaches have been proposed recently for modeling dependency among genes within each gene set. Also, it is worth noting that pairwise information between gene sets is of critical importance and more efforts should be addressed to develop new methodologies.

Acknowledgements

The authors are grateful to the referees for their constructive comments which significantly improved this paper. The authors are grateful to the Ministry of Science and Technology, R. O. C. for funding support (MOST 104-2118-M-002-004).

References

Brannon AR, Haake SM, Hacker KE, Pruthi RS, Wallen EM, Nielsen ME, et al. (2012). Meta-analysis of clear cell renal cell carcinoma gene expression defines a variant subgroup and identifies gender influences on tumor biology. *European Urology*, 61(2): 258–268.

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11(94): 1–13.

Law CW, Chen Y, Shi W, Smyth GK (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2): R29.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7): 621–628.

Nam D (2010). De-correlating expression in Gene-set analysis. *Bioinformatics*, 26(18): i511–i516.

Rahmatallah Y, Emmert-Streib F, Glazko G (2012). Gene set analysis for self-contained tests: Complex null and specific alternative hypotheses. *Bioinformatics*, 28(23): 3073–3080.

Ren X, Hu Q, Liu S, Wang J, Miecznikowski JC (2017). Gene set analysis controlling for length bias in RNA-seq experiments. *BioData Mining*, 10(1): 5.

Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3): R25.

Schäfer J, Strimmer K (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting Genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550.

Székely GJ, Rizzo ML, et al. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10): 1249–1272.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38): 13544–13549.

Wang X, Cairns MJ (2013). Gene set enrichment analysis of RNA-seq data: Integrating differential expression and splicing. In: *BMC Bioinformatics*, volume 14, S16. BioMed Central.

Wang Z, Gerstein M, Snyder M (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1): 57–63.

Yaari G, Bolen CR, Thakar J, Kleinstein SH (2013). Quantitative set analysis for Gene expression: A method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Research*, 41(18): e170.

Zhang W, Huang RS, Duan S, Dolan ME (2009). Gene set enrichment analyses revealed differences in gene expression patterns between males and females. *In Silico Biology*, 9(3): 55–63.

Zhang W, Long H, He B, Yang J (2018). DECtp: Calling differential gene expression between cancer and normal samples by integrating tumor purity information. *Frontiers in Genetics*, 9: 321.