12-2017

Sample size calculations and normalization methods for RNA-seq data.

Xiaohong Li
University of Louisville

SAMPLE SIZE CALCULATIONS AND NORMALIZATION METHODS FOR RNA-SEQ DATA

By

Xiaohong Li
B.S. Cleveland State University, 2002
M.S. University of Louisville, 2010

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
Of the University of Louisville
In Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

December 2017

SAMPLE SIZE CALCULATIONS AND NORMALIZATION METHODS FOR RNA-SEQ DATA

By
Xiaohong Li
B.S. Cleveland State University, 2002
M.S. University of Louisville, 2010

A Dissertation Approved on

July 17, 2017

by the following Dissertation Committee:

Shesh N. Rai, Ph.D. Adviser

Dongfeng Wu, Ph.D. Co-Adviser

Guy N. Brock, Ph.D. Co-Adviser

Eric C. Rouchka, Ph.D.

Ryan S. Gill, Ph.D.

Jeremy Gaskins, Ph.D.

ACKNOWLEDGMENTS

I would like to thank my advisor and mentor, Dr. Shesh N. Rai, for his support, patience and guidance. His unique teaching style has encouraged my growth and confidence in becoming an independent researcher in the Bioinformatical and Statistical fields. I would also like to thank the committee co-advisers, Dr. Dongfeng Wu and Dr. Guy N. Brock, and committee members, Dr. Eric C. Rouchka, Dr. Ryan S. Gill and Dr. Jeremy Gaskins for their support, assistance and comments. I would also like to thank Dr. Nigel G.F. Cooper for his support and encouragement as a boss and adviser. No words can describe how thankful I feel to work for such a wonderful employer. I would also like to express my thanks to my lovely husband, Dr. Timothy E. O'Toole, for standing by me with tremendous patience and support. I would also like to thank Dr. Craig Blakely and Dr. Jack Barnette for their advice, understanding and support. I would like to thank Dr. Rudy Parrish, Dr. Alexander Cambon, Patrick Trainer, Jasmit Shah and Whitney T. Rogers for their support. I would like to thank all of my coworkers in KBRIN Bioinformatics core, KBRIN administrative office and Genomics core for their support. Finally, I would like to thank my lovely parents and my three wonderful children for their understanding and support.

ABSTRACT

SAMPLE SIZE CALCULATIONS AND NORMALIZATION METHODS FOR RNA-SEQ DATA

Xiaohong Li

July 17, 2017

High-throughput RNA sequencing (RNA-seq) has become the preferred choice for transcriptomics and gene expression studies. With the rapid growth of RNA-seq applications, sample size calculation methods for RNA-seq experiment design and data normalization methods for DEG analysis are important issues to be explored and discussed. The underlying theme of this dissertation is to develop novel sample size calculation methods in RNA-seq experiment design using test statistics. I have also proposed two novel normalization methods for analysis of RNA-seq data. In chapter one, I present the test statistical methods including Wald's test, log-transformed Wald's test and likelihood ratio test statistics for RNA-seq data with a negative binomial distribution. Following the test statistics, I present the five sample calculation methods based on a one-sided test. A comparison of my five methods and an existing method was performed by calculating the sample sizes and the simulated power in different scenarios. Due to the limitations of these methods, in chapter two, I have further derived two explicit sample size calculation methods based on a generalized linear model with a negative binomial distribution in RNA-seq data. These two sample size methods based on a two-sided Wald's test are presented under a wide range of settings including the imbalanced design and unequal read depth, which is applicable in many situations. In chapter 3, I have a literature review of the existing normalization methods and describe the challenge of choosing an optimal normalization method due to multiple factors contributing to read count variability that effect overall the sensitivity and specificity. Then, I present two proposed normalization methods. I evaluate the performance of the commonly used methods (DESeq, TMM-edgeR, FPKM-CuffDiff, TC, Med, UQ and FQ) and two new methods I propose: Med-pgQ2 and UQ-pgQ2. The results from MAQC2 data shows that my proposed Med-pgQ2 and UQ-pgQ2 methods may be better choices for the differential gene analysis of RNA-seq data by improving specificity while maintaining a good detection power given a nominal FDR level. Finally, in chapter 4, I focus on data analysis in RNA-seq data using three normalization methods and two test statistic method with the aid of *DESeq2* and *edgeR* packages. Through within-group analysis of these real RNA-seq data, I have found my normalization method, UQ-pgQ2, performs best with a lower false positive rate while maintaining a good detection power. Thus, in my work, I have derived the explicit sample size calculation methods, which is a very useful tool for researchers to quickly estimate the sample sizes in an experiment design. Furthermore, my two normalization methods can improve the performance for differential gene analysis of RNA-seq data by controlling false positives for high read count genes.

Key Words: Sample sizes; RNA-seq; Normalization methods; Power; Differentially Expressed Genes (DEGs).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

SAMPLE SIZE CALCULATION METHODS BASED ON STATISTICAL TESTS IN A NEGATIVE BINOMIAL DISTRIBUTION FOR RNA-SEQ DATA

1. Introduction

Sample size calculations are a prerequisite in an experimental design for biological research and clinical trials. Recently, high-throughput RNA sequencing (RNA-seq) technology has been widely used for gene expression studies in a variety of applications, such as expression profiling of mRNAs or non-coding RNA [1-3], *de novo* assembly and characterization of transcriptomes [4, 5], and the identification of novel alternatively spliced transcripts [6, 7]. These novel transcripts or differentially expressed genes (DEGs) identified from RNA-seq data may serve as human disease biomarkers or gene signatures for the clinical diagnosis [8-10]. With the rapid growth of RNA-seq applications, sample size calculation methods derived from test statistics with an appropriate distribution are important issues to be explored.

Due to the initial high cost of RNA-seq, sample size, in terms of the number of biological replicates, was not seriously considered as part of the experimental design. As a case in point, one RNA-seq review article [11] documented several RNA-seq studies showing that many had only one or a few biological replicates. While thousands of DEGs were identified within these studies, the lack of biological replicates leads to an absence of knowledge concerning biological variations and may result in a high percentage of false positive genes. Therefore, ignorance of biological variation is the fundamental problem with the analyzed results collected from un-replicated data. A recent paper [12] was the first to point out that conclusions drawn from un-replicated samples can be misleading and unrealistic. Later, another study [13] further addressed design and validation issues due to the lack of biological replicates in RNA-seq data.

One of the key questions in an experimental design is to determine the number of biological replicates needed for differential expression analysis in order to achieve a desired statistical power given a significance level α and an underlying distribution. Since RNA-seq data are read counts, a Poisson distribution is commonly used as the model for identifying DEGs in RNA-seq data [14, 15]. Fang and Cui (2011) were the first one to derive the sample size calculations based on a Wald statistical test with a Poisson distribution for single gene in RNA-seq. Later, several sample size calculation methods that were derived from the score statistic and the log-likelihood ratio test (LRT) statistic using the Poisson distribution were proposed [16]. However, the assumption of a Poisson distribution that the expected mean and variance are equal usually does not hold for RNA-seq studies, where the variance is typically greater than the mean of the read counts [17]. Therefore, a negative binomial distribution with a dispersion parameter is used to model RNA-seq data by the existing software packages such as *DESeq* [17] and *edgeR* [18], in which an exact test is used to test DEGs between conditions. Subsequently, a sample size calculation method based on an exact test statistic with the aid of the *edgeR* package [18] was proposed [19]. However, sample size methods derived from other test statistics such as the Wald test, the LRT and an extension of Wald test via log-transformation using negative binomial distribution to model the RNA-seq data have not yet been explored.

Like microarray data, an RNA-seq dataset contains thousands of genes to be tested simultaneously and independently for differential expression analysis. A method for the adjustment or correction of p-values is required to control the type I error rate when multiple pairwise comparisons are performed. Instead of setting the critical value $\boldsymbol{\alpha}$ at 0.05 or 0.01 for significance, a much lower critical value $\boldsymbol{\alpha}^*$ is required to correct for the *inflation* of $\boldsymbol{\alpha}$. The most common method to control the family-wise error rate (FWER) is the Bonferroni correction in which the adjusted p-value is computed via dividing the critical p-value by the total number of comparisons being made. The other widely used method for this multiple correction problem is an FDR correction [20]. Since the Bonferroni correction with a large number of tests is more conservative than the FDR correction, using the Bonferroni correction results in a cost of increasing the probability of producing false negatives and consequently reducing the statistical power. For high dimensional microarray data analysis, an extension of the FDR correction was proposed [21] and is widely used by many researchers. To address similar issues for high-dimensional RNA-seq data, a sample size determination based on the extended FDR correction from microarray data analysis was further proposed [16].

Our study is motivated by exploring sample size calculations using the well-known test statistics (the Wald test and LRT) and a negative binomial distribution to model RNA-seq data. In Section 2, we first define the Wald test, the log-transformed Wald test and the LRT statistics using the negative binomial distribution to model a single gene in RNA-seq data. Then, we derive sample size calculations based on these defined statistical tests. Lastly, we derived sample size calculation methods for testing multiple genes while controlling the FDR [16]. In Section 3, we simulated power for testing single gene and multiple genes corresponding to the sample sizes estimated from our proposed methods and an existing method. The performance of these six methods is compared and evaluated via the required sample sizes and the estimated power. An application of real RNA-seq data to illustrate sample size calculations is presented in Section 4. Finally, we end with a discussion and conclusion in Section 5.

2. Methods

Derivation of Test Statistics

In an RNA-seq experiment, the data contains thousands of genes $(g = 1, \dots, G)$ with different number of reads for each sample mapped to the reference genome. Since the total number of reads among samples is different, the distribution of the gene in the sample with the same condition is not identical. A normalization factor called the size factor is used to model RNA-seq data with a negative binomial distribution. For simplicity, the following statistical tests and consequent sample size calculation methods are based on a single gene tests for DEG analysis.

For a single gene in RNA-seq data, suppose that, for each condition i $(i = 0, 1)$, the observations X_{ij} ($j = 1, \dots, n_i$), are independent and identically follow a negative binomial distribution as $X_{ij} \sim NB(s_{ij}\gamma_i, \phi)$ [17, 22]. Under this setting, γ_i is the true gene expression level in condition i, s_{ij} is a size factor to normalize the raw read for the total number of reads mapped in the sample j, and ϕ is a dispersion parameter with the assumption $\phi_i = \phi$. Thus, the summation of reads per gene per condition ($X_i =$

$\sum_{j=1}^{n_i} X_{ij}$) also follows a negative binomial distribution with parameters $nu_i = t_i = s_i\gamma_i$ and ϕ /*n*, where $s_i = \sum_{j=1}^{n_i} s_{ij}$ is the summation of the size factor for mapping reads in condition *i* and *n* is the number of biological replicates with the assumption $n_i = n$.

For detection of a differentially expressed gene from RNA-seq data, the ratio γ_1/γ_0 typically represents the fold change ($\rho = \gamma_1/\gamma_0$). If the fold change equals one, we can say that this gene is not differentially expressed. Therefore, we are interested in making an inference about the ratio using the Wald statistics and the likelihood ratio methods for sample size calculations.

For testing the hypothesis about the fold change in regards to the DEGs, it is equivalent to test the hypothesis,

$$H_0: \gamma_1 = \gamma_0 \text{ vs. } H_1: \gamma_1 \neq \gamma_0.$$

Since γ_i $(i = 0, 1)$ and ϕ are unknown parameters, we use the following two sample estimates for these parameters under the negative binomial distribution in RNA-seq data [23].

Unconstrained Maximum Likelihood Estimate (MLE): The likelihood and log-likelihood function of $X_0 \sim NB(s_0\gamma_0, \frac{\phi}{n})$ and $X_1 \sim NB(s_1\gamma_1, \frac{\phi}{n})$ are:

$$\mathrm{L}(\gamma_0, \gamma_1, \phi | x_0, x_1) = \prod_{i=0}^{1} \frac{\Gamma\left(\frac{n}{\phi} + x_i\right)}{\Gamma\left(\frac{n}{\phi}\right) \cdot x_i!} \left(\frac{s_i\gamma_i\phi/n}{(s_i\gamma_i\phi/n)+1}\right)^{x_i} \left(\frac{1}{(s_i\gamma_i\phi/n)+1}\right)^{\frac{n}{\phi}},$$

$$\ln \mathrm{L}(\gamma_0, \gamma_1, \phi | x_0, x_1) = \sum_{i=0}^{1} \ln \Gamma\left(\frac{n}{\phi} + x_i\right) - 2\ln\Gamma\left(\frac{n}{\phi}\right) - \sum_{i=0}^{1} \ln x_i! + \sum_{i=0}^{1} x_i \ln(s_i) +$$

$$\sum_{i=0}^{1} x_i \ln(\gamma_i) + \sum_{i=0}^{1} x_i \ln\left(\frac{\phi}{n}\right) - \sum_{i=0}^{1} \left(x_i + \frac{n}{\phi}\right) \ln\left(\frac{s_i\gamma_i\phi}{n} + 1\right),$$

where $t_0 = nu_0 = s_0\gamma_0$ and $t_1 = nu_1 = s_1\gamma_1$.

Setting the first derivative to zero, we obtain the unrestricted MLEs with respect to γ_i: $\hat{\gamma}_0 = \frac{x_0}{s_0}$ and $\hat{\gamma}_1 = \frac{x_1}{s_1}$. Subsequently, we can obtain the MLE of $\hat{\phi}$ by solving the following equation:

$$\frac{\partial l}{\partial \phi} = \sum_{i=0}^{1} \psi\left(\frac{n}{\phi} + x_i\right) - 2\psi\left(\frac{n}{\phi}\right) + \frac{n}{\phi^2}\{ln(\bar{x}_0\phi + 1) + ln(\bar{x}_1\phi + 1)\} - \frac{\bar{x}_0\left(x_0 + \frac{n}{\phi}\right)}{(\bar{x}_0\phi + 1)} - \frac{\bar{x}_1\left(x_1 + \frac{n}{\phi}\right)}{(\bar{x}_1\phi + 1)} = 0.$$

Several mathematical optimization methods, such as the Newton-Raphson method, can be used to estimate $\hat{\phi}$ for the above equation. Since there is no closed form to estimate the dispersion ϕ, we derived the sample size formula based on a constant value of ϕ estimated from the data.

Constrained Maximum Likelihood Estimate (CMLE) for γ_i and ϕ: The parameters are estimated under the null hypothesis $H_0: \gamma_0 = \gamma_1$. Let $w = \frac{s_1}{s_0}, \rho = \frac{\gamma_1}{\gamma_0} = \frac{t_1/s_1}{t_0/s_0}$ and $t_1 = \rho w t_0 = \rho w n u_0$. Then, the log likelihood function is re-parameterized as:

$$\ln L(\rho, t_0, \phi | x_0, x_1) = \sum_{i=0}^{1} \ln\Gamma\left(\frac{n}{\phi} + x_i\right) - 2\ln\Gamma\left(\frac{n}{\phi}\right) - \sum_{i=0}^{1} \ln x_i! + x_0 \ln t_0 + x_0 \ln(\phi) + x_0 \ln\left(\frac{1}{n}\right) + x_1 \ln(t_0) + x_1 \ln(w/n) + x_1 \ln(\phi) + x_1 \ln(\rho) - x_0 \ln\left(\frac{t_0\phi}{n} + 1\right) - x_1 \ln\left(\frac{\rho w t_0 \phi}{n} + 1\right) - \frac{n}{\phi}\ln\left(\frac{t_0\phi}{n} + 1\right) - \frac{n}{\phi}\ln\left(\frac{\rho w t_0\phi}{n} + 1\right).$$

Using partial derivatives of this equation with respect to ρ and t_0, the MLEs in the unrestricted parameter space are $\hat{t}_0 = x_0$ and $\hat{\rho} = \frac{x_1}{w x_0}$. However, the CMLE $\tilde{t}_0$ in the constrained parameter space and under $H_0: \rho = 1$ is given by

$$\frac{\partial l(\rho=1,t_0,\phi|x_0,x_1)}{\partial t_0} = \frac{x_0-t_0}{t_0\left(\frac{\phi t_0}{n}+1\right)} + \frac{x_1-wt_0}{t_0\left(\frac{w\phi t_0}{n}+1\right)}.$$

Setting this equation to zero, we obtain

$$\tilde{t}_0 = \frac{\sqrt{[\phi(wx_0+x_1)-n(1+w)]^2+8w\phi n(x_0+x_1)}}{4w\phi} + \frac{[\phi(wx_0+x_1)-n(1+w)]}{4w\phi}.$$

Thus, setting the derivative of ϕ to the zero and $t_0 = \tilde{t}_0$ and $\rho = 1$, the CMLE $\tilde{\phi}$ for ϕ is estimated by solving the following equation:

$$0 = \sum_{i=0}^{1} \Psi\left(\frac{n}{\tilde{\phi}} + x_i\right) - 2\Psi\left(\frac{n}{\tilde{\phi}}\right) + \frac{x_0-\tilde{t}_0}{\tilde{\phi}\left(\frac{\tilde{t}_0\tilde{\phi}}{n}+1\right)} + \frac{x_1-w\tilde{t}_0}{\tilde{\phi}\left(\frac{w\tilde{t}_0\tilde{\phi}}{n}+1\right)} + \frac{n}{\tilde{\phi}^2}\left\{\ln\left(\frac{\tilde{t}_0\tilde{\phi}}{n}+1\right) + \ln\left(\frac{w\tilde{t}_0\tilde{\phi}}{n}+1\right)\right\}.$$

We finally obtain

$$\tilde{\gamma}_1 = \tilde{\gamma}_0 = \tilde{t}_0/s_0 = \frac{\sqrt{[\phi\,(wx_0+x_1)-n(1+w)]^2+8w\phi\, n(x_0+x_1)}}{4ws_0\phi} + \frac{[\phi\,(wx_0+x_1)-n(1+w)]}{4ws_0\phi},$$

and $\tilde{u}_0 = \frac{\tilde{t}_0}{n} = \frac{\sqrt{[\phi(w\bar{x}_0+\bar{x}_1)-(1+\rho w)]^2+8w\phi(\bar{x}_0+\bar{x}_1)}}{4w\phi} + \frac{[\phi(w\bar{x}_0+\bar{x}_1)-(1+w)]}{4w\phi}.$

Since there is no closed form to estimate the dispersion ϕ from MLE and CMLE, we derived the sample size formula based on a fixed and constant value. For simplicity, we set the dispersion in MLE and CMLE to be equal with a combination of fixed dispersion (0.1, 0.5 and 1).

Wald statistical test and log-transformed Wald statistical test

Wald statistical test: The Wald statistical test is an asymptotic test based on the normal approximation, which utilizes the large-sample properties of the MLE. Following procedures from the studies [24-26] for comparing two independent Poisson rates with unequal sample frames, we derived the Wald's inference procedures using the properties of $\hat{\gamma}_i$ and $\hat{\phi}$ estimated from MLE and $\tilde{\gamma}_i$and $\tilde{\phi}$ from CMLE with a negative binomial distribution in two conditions ($\hat{\gamma}_i = \frac{X_i}{s_i}$), where X_0 and X_1 in two conditions are assumed to be independent. For simplicity, we set $\hat{\phi} = \tilde{\phi} = \phi$ with a constant for the sample size and power analysis.

The null hypothesis H_0 is equivalent to H_0: $\gamma_1 - \gamma_0 = 0$, and consequently we make inferences based on the quantity $T = \hat{\gamma}_1 - \hat{\gamma}_0 = \frac{X_1}{s_1} - \frac{X_0}{s_0}$. In this case, the variance of T is $\sigma_T^2 = \frac{\gamma_1}{s_1} + \frac{\gamma_0}{s_0} + \frac{\phi}{n}(\gamma_1^2 + \gamma_0^2)$ and can be estimated by $s_T^2 = \frac{\hat{\gamma}_1}{s_1} + \frac{\hat{\gamma}_0}{s_0} + \frac{\phi}{n}(\hat{\gamma}_1^2 + \hat{\gamma}_0^2)$, where the parameters are estimated form MLE. Thus, Wald statistical test from MLE can be obtained by the statistic T/S_T

$$Z_{w1} = \frac{X_1 - wX_0}{\sqrt{X_1 + w^2 X_0 + \frac{\hat{\phi}}{n}(X_1^2 + w^2 X_0^2)}}, \tag{1}$$

where $w = \frac{s_1}{s_0}$, is the ratio of total size factors between the two conditions. Similarly, for $T = \frac{X_1}{s_1} - \frac{X_0}{s_0}$, σ_T^2 can be estimated by $s_T^2 = \frac{\tilde{\gamma}_1}{s_1} + \frac{\tilde{\gamma}_0}{s_0} + \frac{\tilde{\phi}}{n}(\tilde{\gamma}_1^2 + \tilde{\gamma}_0^2)$ using the parameters estimated from CMLE. By substituting $\tilde{\gamma}_i = \frac{n\tilde{u}_0}{s_0}$, the 2nd Wald statistical test can be obtained by the statistic T/S_T:

$$Z_{w2} = \frac{X_1 - wX_0}{\sqrt{nw^2\tilde{u}_0(1/w + 1 + 2\tilde{\phi}\tilde{u}_0)}}. \tag{2}$$

Log-transformation of Wald's statistical test*:* To test the null hypothesis, it is also equivalent to test H_0: $\ln\left(\frac{\gamma_1}{\gamma_0}\right) = 0$. The logarithmic transformation is usually adopted for skewness correction and variance stabilization as suggested by these studies [16, 24].

The statistical inference on the quantity is $U = \ln(\hat{\gamma}_1/\hat{\gamma}_0) = \ln\left(\frac{X_1}{s_1}\right) - \ln\left(\frac{X_0}{s_0}\right)$. Since $\frac{X_1}{s_1}$ and $\frac{X_0}{s_0}$ have asymptotically normal distributions, $N(\gamma_1, \frac{\gamma_1}{s_1} + \frac{\phi}{n}\gamma_1^2)$ and $N\left(\gamma_0, \frac{\gamma_0}{s_0} + \frac{\phi}{n}\gamma_0^2\right)$, respectively, $\ln\frac{X_1}{s_1}$ and $\ln\frac{X_0}{s_0}$ correspondingly also have an asymptotically normal $N(\ln\gamma_1, \frac{1}{s_1\gamma_1} + \frac{\phi}{n})$ and $N\left(\ln\gamma_0, \frac{1}{s_0\gamma_0} + \frac{\phi}{n}\right)$ by the Delta Method and Slutsky's theorem. Thus, the variance of U is $\sigma_U^2 = Var(U) = Var\left[\ln\left(\frac{X_1}{s_1}\right) - \ln\left(\frac{X_0}{s_0}\right)\right] = \frac{1}{s_1\gamma_1} + \frac{1}{s_0\gamma_0} + 2\frac{\phi}{n}$. U/S_U can be used for testing H_0, when $s_U^2 = \frac{1}{s_1\hat{\gamma}_1} + \frac{1}{s_0\hat{\gamma}_0} + 2\frac{\hat{\phi}}{n}$ using the parameters estimated from MLEs. Thus, a log-transformation of Z_{w1} is applied and the test statistic is defined as

$$Z_{lw1} = \frac{\ln(\frac{X_1}{X_0}) - \ln w}{\sqrt{1/X_1 + \frac{1}{X_0} + 2\frac{\hat{\phi}}{n}}}, \text{ where } \hat{\gamma}_i = \frac{X_i}{s_i} \text{ and } \phi = \hat{\phi}. \tag{3}$$

Similarly, U/S_U can be used for testing H_0, when $S_U^2 = \frac{1}{s_1\tilde{\gamma}_1} + \frac{1}{s_0\tilde{\gamma}_0} + 2\frac{\phi}{n}$ using the parameters estimated from CMLE and then, we apply the log-transformation of Z_{w2} and use the test statistic

$$Z_{lw2} = \frac{\ln(\frac{X_1}{X_0}) - \ln w}{\sqrt{\frac{1}{n\tilde{u}_0}\left(\frac{1}{w} + 1 + 2\tilde{u}_0\tilde{\phi}\right)}}, \text{ where } \tilde{\gamma}_1 = \tilde{\gamma}_0 = \frac{\tilde{t}_0}{s_0} = \frac{n\tilde{u}_0}{s_0}. \tag{4}$$

We note that the equations defined in (3) and (4) do not exist when $X_0 = 0$ or $X_1 = 0$. In this case, X_0 or X_1 was adjusted to 0.5 [24, 25].

Generalized Likelihood Ratio Test (GLRT)

The GLRT statistic is defined as the ratio of the maximum value of the likelihood function under the restriction of the null hypothesis to the maximum likelihood function under the unrestricted parameter space. For a vector of parameters $\theta \in \boldsymbol{\Theta}$, the GLRT for H_0: $\theta \in \boldsymbol{\Theta_0}$ versus H_1: $\theta \in \boldsymbol{\Theta_1}$ is expressed as

$$\lambda\{x_0, x_1 | \theta(\gamma_i, \phi)\} = \frac{\sup\{L(\theta(\gamma_i,\phi)|x_0,x_1):\theta\in\boldsymbol{\Theta_0}\}}{\sup\{L(\theta(\gamma_i,\phi)|x_0,x_1):\theta\in\boldsymbol{\Theta}\}},$$

where $L(\theta(\gamma_i, \phi)|x_0, x_1)$ is the likelihood function defined. The denominator $\sup\{L(\theta(\gamma_i, \phi)|x_0, x_1): \theta \in \Theta\}$ in equation (14) is obtained using the MLE of θ, where $\hat{\gamma}_0 = \frac{X_0}{s_0}$ and $\hat{\gamma}_1 = \frac{X_1}{s_1}$. The numerator $\sup\{L(\theta(\gamma_i, \phi)|x_0, x_1): \theta \in \boldsymbol{\Theta_0}\}$ is obtained using the CMLE of θ under H_0, where $\rho = \frac{\gamma_1}{\gamma_0} = 1$. So the GLRT statistic is defined as

$$Z_{lrt} = \frac{\sup\{L(\theta(\tilde{\gamma}_1=\tilde{\gamma}_0, \tilde{\phi},\rho=1)|x_0,x_1):\theta(\gamma_0,\gamma_1,\phi)\in\boldsymbol{\Theta_0}\}}{\sup\{L(\theta(\hat{\gamma}_0,\hat{\gamma}_1,\hat{\phi})|x_0,x_1):\theta(\gamma_0,\gamma_1,\phi)\in\boldsymbol{\Theta}\}} =$$

$$\frac{\binom{\frac{n}{\tilde{\phi}}+x_0-1}{x_0}\left(\frac{\tilde{u}_0\,\tilde{\phi}}{\tilde{u}_0\,\tilde{\phi}+1}\right)^{x_0}\left(\frac{1}{\tilde{u}_0\,\tilde{\phi}+1}\right)^{\frac{n}{\tilde{\phi}}}\times\binom{\frac{n}{\tilde{\phi}}+x_1-1}{x_1}\left(\frac{w\tilde{u}_0\,\tilde{\phi}}{w\tilde{u}_0\,\tilde{\phi}+1}\right)^{x_1}\left(\frac{1}{w\tilde{u}_0\,\tilde{\phi}+1}\right)^{\frac{n}{\tilde{\phi}}}}{\binom{\frac{n}{\tilde{\phi}}+x_0-1}{x_0}\left(\frac{x_0\hat{\phi}/n}{x_0\hat{\phi}/n+1}\right)^{x_0}\left(\frac{1}{\frac{x_0\hat{\phi}}{n}+1}\right)^{\frac{n}{\tilde{\phi}}}\times\binom{\frac{n}{\tilde{\phi}}+x_1-1}{x_1}\left(\frac{x_1\hat{\phi}/n}{x_1\hat{\phi}/n+1}\right)^{x_1}\left(\frac{1}{\frac{x_1\hat{\phi}}{n}+1}\right)^{\frac{n}{\tilde{\phi}}}}.$$

Since $T = -2\ln Z_{lrt} = -2\left[\ln L(\tilde{\gamma}_1 = \tilde{\gamma}_0, \tilde{\phi}, \rho = 1) - \ln L(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\phi})\right]$ approximately follows a χ_1^2 distribution, the p-value for the one-sided test statistic of T is approximately [25]:

$$p.value(x_0, x_1) = 0.5\{1 - \chi_1^2(T)\}. \tag{5}$$

The p-value in equation (5) is further adjusted by the FDR correction when multiple genes are used for the data analysis. Combining these together, the parameter estimates based on the MLEs from two assumptions and the following test statistics are summarized in Table 1.

Table 1: Statistical tests are used for deriving sample size calculations. The parameters are estimated using MLEs and CMLE methods.

Statistic tests	**Maximum Likelihood estimates (MLE)**	**Statistical test**	**Log Transformed test**
Wald test	MLE under unrestricted parameter space	Z_{w1}	Z_{lw1}
	Conditional MLE (CMLE) under H_0: $\gamma_0 = \gamma_1$	Z_{w2}	Z_{lw2}
Generalized Likelihood ratio test	CMLE/MLE	Z_{lrt}	

γ_0 and γ_1 are true gene expression between two conditions.

Sample size calculation for a single gene

In order to calculate the sample size, a power function needs to be constructed. The power of a test is the probability that the null hypothesis is rejected when the alternative hypothesis is true. We derive the sample size under the specified power 1-$\boldsymbol{\beta}$ and the significance level $\boldsymbol{\alpha}$ with an assumption of a balanced design experiment between conditions (i.e., $n_0 = n_1 = n$) and one-sided statistical test under H_1: $\frac{\gamma_1}{\gamma_0} = \rho > 1$.

Derivation of sample size based on the Wald test statistics: Under the null hypothesis $H_0: \gamma_1 = \gamma_0$, Z_w has asymptotically standard normal distribution. Thus, at type I error rate α, H_0 is rejected when $|Z_{w1}(x_1, x_0)| > z_{1-\alpha/2}$, where $Z_w(x_1, x_0)$ is the observed value of Z_w and $z_{1-\alpha/2}$ is the standard normal distribution. For a specific alternative hypothesis $H_1: \frac{\gamma_1}{\gamma_0} = \rho > 1$, the power of the one-sided test is

$$\Pr\left(Z_w > z_{1-\alpha} | \gamma_1 = \rho\gamma_0, \rho > 1\right). \tag{6}$$

Sample size calculation of $\mathbf{Z_{w1}}$ from MLE*:* Under the null hypothesis, Z_w has an asymptotically standard normal distribution. From equation (1), the critical region for Z_{w1} at the α significance level consists of those points (X_1, X_0) that satisfy the inequality (6) is

$$X_1 - wX_0 > z_{1-\alpha}\sqrt{X_1 + w^2X_0 + \frac{\hat{\phi}}{n}(X_1^2 + w^2X_0^2)}.$$

Under H_1, $X_1 - wX_0$ in the above equation is asymptotically normal with mean u_{w1} and variance σ_{w1}^2 given by

$u_{w1} = E(X_1 - wX_0) = wnu_0(\rho - 1)$ and

$\sigma_{w1}^2 = Var(X_1 - wX_0) = wnu_0(\rho + w + \phi w\rho^2 u_0 + \phi wu_0),$

where $= \frac{s_1}{s_0}$, $\rho = \frac{\gamma_1}{\gamma_0}$ and $s_0\gamma_0 = n\bar{s}_0\gamma_0 = nu_0$ under the assumption where X_0 and X_1 are independent. The expression $\sqrt{X_1 + w^2X_0 + \frac{\hat{\phi}}{n}(X_1^2 + w^2X_0^2)}$ converges in probability to $\sqrt{wnu_0(\rho + w + \phi w\rho^2 u_0 + \phi wu_0)}$ $(\hat{\gamma}_i = \frac{X_i}{s_i}, \hat{\phi} = \phi,\ X_i \xrightarrow{p} s_i\gamma_i$ and $X_i^2 \xrightarrow{p} (s_i\gamma_i)^2$.

Hence, the approximate power *Pr* can be expressed in the terms of the cumulative normal distribution as:

$$\Pr(Z_{w1} > z_{1-\alpha} | \gamma_1 = \rho\gamma_0, \rho > 1) = \Phi\left[z_{1-\alpha} - (\rho - 1)\sqrt{\frac{nu_0}{(1+\rho/w+\phi\rho^2u_0+\phi u_0)}}\right],$$

where $\Phi(\cdot)$ is the standard normal distribution function. Let the power $Pr = 1 - \beta$**,** we can obtain:

$$1 - \beta = 1 - \Phi\left[z_{1-\alpha} - (\rho - 1)\sqrt{\frac{nu_0}{\left(1+\frac{\rho}{w}+\phi\rho^2u_0+\phi u_0\right)}}\right],$$

and solving the equation, we can show

$$n_{w1} = \frac{(1+\rho/w+\phi\rho^2u_0+\phi u_0)(z_{1-\alpha}+z_{1-\beta})^2}{u_0(\rho-1)^2}, \tag{7}$$

where ϕ is assumed to be a constant value, ϕ and u_0 can be estimated from RNA-seq samples. When the alternative hypothesis is $H_1: \frac{\gamma_1}{\gamma_0} = \rho < 1$, equation (7) is also true.

Sample size calculation of $\mathbf{Z_{lw1}}$: Under the null hypothesis, Z_{lw1} has an asymptotically standard normal distribution. The critical region for Z_{lw1} from equation (3) at the α significance level consists of those points (X_1, X_0) that satisfy the inequality (6) is

$$\ln(\frac{X_1}{X_0}) - \ln w > z_{1-\alpha}\sqrt{1/X_1 + \frac{1}{X_0} + 2\frac{\hat{\phi}}{n}}.$$

Under H_1, $\ln(\frac{X_1}{X_0}) - \ln w$ in the above equation is asymptotically normal with mean u_{lw1} and variance σ_{lw1}^2 given by

$$u_{lw1} = E\left(\ln\left(\frac{X_1}{X_0}\right) - \ln w\right) = \ln \gamma_1 - \ln \gamma_0 = \ln \rho \text{ and}$$

$$\sigma_{lw1}^2 = Var\left(\ln \frac{X_1}{s_1} - \ln \frac{X_0}{s_0}\right) = \frac{1}{s_0\gamma_0 w\rho} + \frac{1}{s_0\gamma_0} + 2\frac{\phi}{n} = \frac{1}{nu_0}\left(\frac{1}{w\rho} + 1 + 2\phi u_0\right),$$

where $= \frac{s_1}{s_0}$, $\rho = \frac{\gamma_1}{\gamma_0}$, and $d_0\gamma_0 = nu_0$. The expression $\sqrt{1/X_1 + \frac{1}{X_0} + 2\frac{\hat{\phi}}{n}}$ has asymptotic limit $\sqrt{1/d_1\gamma_1 + \frac{1}{d_0\gamma_0} + 2\frac{\phi}{n}} = \sqrt{\frac{1}{nu_0}\left(\frac{1}{w\rho} + 1 + 2\phi u_0\right)}$. Hence, the approximate power Pr can be expressed in the terms of the cumulative normal distribution as:

$$\Pr\left(Z_{lw1} > z_{1-\alpha} | \gamma_1 = \rho\gamma_0, \rho > 1\right) = \Phi\left[z_{1-\alpha} - \ln\rho\sqrt{\frac{nu_0}{\left(1 + \frac{1}{w\rho} + 2\phi u_0\right)}}\right],$$

where $\Phi(\cdot)$ is the standard normal distribution function of the standard normal distribution. Let the power $Pr = 1 - \beta$, we can obtain:

$$1 - \beta = 1 - \Phi\left[z_{1-\alpha} - \ln\rho\sqrt{\frac{nu_0}{\left(1+\frac{1}{w\rho}+2\hat{\phi}u_0\right)}}\right],$$

and solving this equation , we can show

$$n_{lw1} = \frac{\left(1+\frac{1}{\rho w}+2\phi u_0\right)\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{u_0(\ln\rho)^2}. \tag{8}$$

Sample size calculation of $\mathbf{Z_{w2}}$ from CMLE: Under the null hypothesis, Z_{w2} has an asymptotically standard normal distribution. The critical region for Z_{w2} from equation (2) at the α significance level consists of those points (X_1, X_0) that satisfy the inequality (6) is

$$X_1 - wX_0 > z_{1-\alpha}\sqrt{nw^2\tilde{u}_0\left(1/w\ +1+2\tilde{\phi}\tilde{u}_0\right)}.$$

Under H_1, $X_1 - wX_0$ in the above equation is asymptotically normal with mean u_{w2} and variance σ_{w2}^2 given by

$$u_{w2} = wnu_0(\rho - 1)\ \text{and}$$

$$\sigma_{w2}^2 = wnu_0(\rho + w + \phi w\rho^2 u_0 + \phi wu_0),$$

where $w = \frac{s_1}{s_0}, \frac{\gamma_1}{\gamma_0} = \ \rho\ and\ \ s_0\gamma_0 = nu_0$. The u_0 estimate from CMLE,

$$\tilde{u}_0 = \frac{\sqrt{[\phi(w\bar{x}_0+\bar{x}_1)-(1+\rho w)]^2+8w\phi(\bar{x}_0+\bar{x}_1)}}{4w\phi} + \frac{[\phi(w\bar{x}_0+\bar{x}_1)-(1+w)]}{4w\phi}, \text{ has limit}$$

$$\frac{\sqrt{[\phi wu_0(1+\rho)-(1+w)]^2+8wu_0\phi(1+\rho w)}}{4w\phi} + \frac{[\phi wu_0(1+\rho)-(1+w)]}{4w\phi}, \tag{9}$$

where $\hat{\gamma}_i = \frac{X_i}{s_i}$, $X_i \xrightarrow{p} s_i\gamma_i, \bar{X}_0 \xrightarrow{p} u_0$ and $\bar{X}_1 \xrightarrow{p} \rho w u_0$. Hence, the approximate power Pr can be expressed in the terms of the cumulative normal distribution as:

$$\Pr\left(Z_{w2} > z_{1-\alpha} | \gamma_1 = \rho\gamma_0, \rho > 1\right)$$

$$= \Phi\left[z_{1-\alpha}\sqrt{\frac{\tilde{u}_0(1/w+1+2\tilde{\phi}\tilde{u}_0)}{u_0(\rho/w+1+\phi\rho^2u_0+\phi u_0)}} - (\rho-1)\sqrt{\frac{nu_0}{(1+\rho/w+\phi\rho^2u_0+\phi u_0)}}\right],$$

where $\Phi(\cdot)$ is the standard normal distribution function of the standard normal distribution. Let the power $Pr = 1 - \beta$, we can obtain:

$$1-\beta = 1-\Phi\left[z_{1-\alpha}\sqrt{\frac{\tilde{u}_0\left(\frac{1}{w}+1+2\tilde{\phi}\tilde{u}_0\right)}{u_0\left(\frac{\rho}{w}+1+\phi\rho^2u_0+\phi u_0\right)}} - (\rho-1)\sqrt{\frac{nu_0}{\left(1+\frac{\rho}{w}+\phi\rho^2u_0+\phi u_0\right)}}\right],$$

and solving this equation, we can show

$$n_{w2} = \frac{(1+\rho/w+\phi\rho^2u_0+\phi u_0)\left(z_{1-\alpha}\sqrt{\frac{\tilde{u}_0(1/w+1+2\phi\tilde{u}_0)}{u_0(\rho/w+1+\phi\rho^2u_0+\phi u_0)}}+z_{1-\beta}\right)^2}{u_0(\rho-1)^2}, \qquad (10)$$

where $\tilde{u}_0$ estimated from CMLE in equation (S.9), $\hat{\phi} = \phi$ assumed as a constant value. When the alternative hypothesis is $H_1\colon \frac{\gamma_1}{\gamma_0} = \rho < 1$, equation (10) is also true.

Sample size calculation of $\mathbf{Z_{lw2}}$ from CMLE: Similarly Z_{lw2} is asymptotically standard normal distribution. The critical region for Z_{lw2} from equation (13) at the α significance level consists of those points (X_1, X_0) is

$$\ln\left(\frac{X_1}{X_0}\right) - \ln w > z_{1-\alpha}\sqrt{\frac{1}{n\tilde{u}_0}\left(\frac{1}{w}+1+2\tilde{u}_0\tilde{\Phi}\right)}.$$

Under H_1, $\ln(\frac{X_1}{X_0}) - \ln w$ in the above equation is asymptotically normal with mean u_{lw2} and variance σ_{lw2}^2 given by

$$u_{lw2} = E\left(\ln\left(\frac{X_1}{X_0}\right) - \ln w\right) = \ln\rho,$$

$$\sigma_{lw2}^2 = \mathrm{Var}\left(\ln\frac{X_1}{s_1} - \ln\frac{X_0}{s_0}\right) = \frac{1}{nu_0}\left(\frac{1}{w\rho} + 1 + 2\phi u_0\right),$$

where $= \frac{s_1}{s_0}$, $\frac{\gamma_1}{\gamma_0} = \rho$, and $d_0\gamma_0 = nu_0$. Hence, the approximated power Pr can be expressed in the terms of the cumulative normal distribution as:

$$\Pr\left(Z_{lw2} > z_{1-\alpha} \mid \gamma_1 = \rho\gamma_0, \rho > 1\right)$$

$$= \Phi\left[z_{1-\alpha}\sqrt{\frac{u_0\left(\frac{1}{w}+1+2\tilde{u}_0\tilde{\phi}\right)}{\tilde{u}_0\left(\frac{1}{w\rho}+1+2u_0\phi\right)}} - \ln\rho\sqrt{\frac{nu_0}{\left(\frac{1}{w\rho}+1+2u_0\phi\right)}}\right],$$

where $\Phi(\cdot)$ is the standard normal distribution function of the standard normal distribution. Let the power $Pr = 1 - \beta$, we can obtain:

$$1 - \beta = 1 - \Phi\left[z_{1-\alpha}\sqrt{\frac{u_0\left(\frac{1}{w}+1+2\tilde{u}_0\hat{\phi}\right)}{\tilde{u}_0\left(\frac{1}{w\rho}+1+2u_0\phi\right)}} - \ln\rho\sqrt{\frac{nu_0}{\left(\frac{1}{w\rho}+1+2u_0\phi\right)}}\right],$$

and solving this equation , we can show

$$n_{lw2} = \frac{\left(\frac{1}{w\rho}+1+2u_0\phi\right)\left(z_{1-\alpha}\sqrt{\frac{u_0\left(\frac{1}{w}+1+2\tilde{u}_0\phi\right)}{\tilde{u}_0\left(\frac{1}{w\rho}+1+2u_0\phi\right)}}+z_{1-\beta}\right)^2}{u_0(\ln\rho)^2}, \tag{11}$$

where $\tilde{u}_0 = \frac{\sqrt{[\phi wu_0(1+\rho)-(1+w)]^2+8wu_0\phi(1+\rho w)}}{4w\phi} + \frac{[\phi wu_0(1+\rho)-(1+w)]}{4w\phi}$, estimated from CMLE under H_0 , and u_0 is the mean read counts under H_1 or the assumed true mean read counts in the control condition. Under the alternative hypothesis H_1: $\frac{\gamma_1}{\gamma_0} = \rho < 1$, equations (7-8, 10-11) are also true.

Derivation of sample size based on the likelihood ratio test (LRT) statistic: For the LRT with a negative binomial distribution, it is difficult to derive a closed-form expression of the power function. We used the method [19] to calculate the power of the LRT given a p-value from the equation (5) under LRT. This method originally borrowed a concept from this study [27] to calculate the power. Given a p-value based on the observed joint probability $P(X_0 = x_0, X_1 = x_1)$, the power under the assumption can be expressed as

$$Pr(n, \rho, u_0, \phi, w, \alpha)$$
$$= \sum_{x_0=0}^{\infty} \sum_{x_1=0}^{\infty} f\left(nw\rho u_0, \frac{\phi}{n}\right) f\left(nu_0, \frac{\phi}{n}\right) I(P(X_0 = x_0, X_1 = x_1)$$
$$< \alpha),$$

where X_0 and X_1 are independent, $f(u, \phi)$ is the probability mass function of the negative binomial distribution with mean u and dispersion ϕ, α is the level of significance, and $I(\cdot)$ is the indicator function of p-value. Thus, given a nominal power $1 - \beta$**,** the power of the test can be represented as the function of the sample size *n* in the form of

$$1 - \beta = Pr(n, \rho, u_0, \phi, w, \alpha). \tag{12}$$

Therefore, the required sample size n to attain the nominal power $1 - \beta$ at a significance level α can then be computed by solving (12) through a numerical approach with respect to *n*.

Sample size calculation with controlling FDR for testing multiple genes

In an RNA-seq experiment, thousands of genes need to be tested simultaneously for DEGs between conditions. In this case, the sample size calculation for a single gene derived above needs to be further adjusted due to the multiple testing problems. In this section we derive sample size calculations by incorporating FDR controlling based on the statistical tests described in the previous sections. The details of controlling FDR have been given in the study [19]. Briefly, FDR (*f*) is defined as

$$f = \frac{m_0\alpha}{m_0\alpha + t_1}, \tag{13}$$

where m_0 is the number of true null hypotheses, $t_1 = E(M_1)$ is the expected number of true rejections, M_0 is the number of false discoveries, M is the total number of genes declared significant, $M_1 = M - M_0$ and f is the control FDR at a specified level. By solving equation (13) with respect to α, the marginal type I error level α^* for the expected number of true rejections t_1 at a given FDR (f) is

$$\alpha^* = \frac{t_1 f}{m_0(1-f)}. \tag{14}$$

Replacing α with α^* in (7-8,10-11), the corresponding sample size calculation formulas corrected by FDR at level f are, respectively,

$$n_{w1} = \frac{\left(1+\rho/w+\hat{\phi}\rho^2 u_0+\hat{\phi} u_0\right)\left(z_{1-\alpha^*}+z_{1-\beta}\right)^2}{u_0(\rho-1)^2}, \tag{15}$$

$$n_{lw1} = \frac{\left(1+\frac{1}{\rho w}+2\hat{\phi} u_0\right)\left(z_{1-\alpha^*}+z_{1-\beta}\right)^2}{u_0(\ln\rho)^2}, \tag{16}$$

$$n_{w2} = \frac{\left(1+\rho/w+\tilde{\phi}\rho^2 u_0+\tilde{\phi}u_0\right)\left(z_{1-\alpha^*}\sqrt{\frac{\tilde{u}_0(1/w+1+2\tilde{\phi}\tilde{u}_0)}{u_0(\rho/w+1+\tilde{\phi}\rho^2 u_0+\tilde{\phi}u_0)}}+z_{1-\beta}\right)^2}{u_0(\rho-1)^2}, \tag{17}$$

$$n_{lw2} = \frac{\left(\frac{1}{w\rho}+1+2u_0\tilde{\phi}\right)\left(z_{1-\alpha^*}\sqrt{\frac{u_0\left(\frac{1}{w}+1+2\tilde{u}_0\tilde{\phi}\right)}{\tilde{u}_0\left(\frac{1}{w\rho}+1+2u_0\tilde{\phi}\right)}}+z_{1-\beta}\right)^2}{u_0(\ln\rho)^2}. \tag{18}$$

Similarly, replacing α with α^* for the LRT statistic, we obtain the function with respect to *n* as

$$1-\beta = Pr(n, \rho, u_0, \phi, w, \alpha^*). \tag{19}$$

Thus, by solving (19) via a numerical approach, the sample size for controlling FDR at level *f* can be obtained.

3. Simulation studies and comparison of results

The proposed sample size formulas are derived from the likelihood function based on large sample theory with an approximate normal distribution. The simulation studies include two parts. In the first part, we calculated sample size based on testing single gene from the different formulas. In the second part, we calculated sample size based on testing the multiple genes using FDR adjusted significance α^* level. The parameter settings in our simulation studies are based on empirical data sets. A comparison of the simulated power for the sample sizes using different methods was performed.

Sample size calculations and power estimation based on testing a single gene

The purpose of this study is to compare the performance of sample size calculations with the estimated power from our formula with the method based on the exact test in the public study [19]. We set the following inputs based on a single gene. Let the type I error rate $\alpha = 0.05$, the power $1 - \beta = 0.8$, the ratio of total size factors between two condition $w = 1$ and 1.2, the mean counts of gene g in control condition $u_0 = 1, 5$ or 10, the dispersion $\phi = 0.1, 0.5$ or 1, and the fold changes $\rho = 1.5, 2, 3$ or 4. Since the read depth across samples in RNA-seq data is usually close to each other, we choose $w = 1.2$ instead of $w = 2$ [19]. For each combination of these designed settings, at first, we used our derived formulas in equations (7-8 and 10-11) to calculate the required sample size, respectively. Then, we calculated the sample size based on the exact test computed using the R codes with the same input settings. Moreover, for each designed setting, we generated 5000 simulations from independent negative binomial distributions based on the calculated sample size *n* given the dispersion ϕ with different mean counts. For the control condition ($i = 0$), we used R to generate random samples given the mean u_0 and ϕ. For the treatment condition ($i = 1$), we generated random samples given mean $u_1 = w\rho u_0$ and ϕ. The test statistics (1-4 and 5) were applied to each simulation sample and the empirical power was obtained as the proportion of simulation samples for which H_0 is rejected with the nominal type I error $\alpha = 0.05$. The results are shown in Table 2, which reports the estimated sample size with associated empirical power given in parentheses under the case w=1 and 1.2.

Table 2: Simulated power corresponding to the sample size for testing a single gene from six methods given the predefined nominal $\alpha = 0.05$ and 80% power.

w	ρ	ϕ	u_0	n_{w1}	n_{w2}	n_{lw1}	n_{lw2}	n_{lrt}	n_{exact}
1	1.5	.1	1	70(.81)	70(.81)	70(.81)	69(.81)	73(.82)	74(.80)
			5	20(.81)	20(.81)	20(.81)	20(.82)	21(.83)	21(.82)
			10	14(.82)	14(.82)	14(.82)	14(.82)	15(.84)	15(.83)
		0.5	1	102(.81)	101(.80)	100(.80)	99(.80)	105(.80)	104(.79)
			5	53(.82)	52(.81)	51(.81)	50(.79)	54(.83)	52(.81)
			10	46(.82)	45(.81)	44(.80)	44(.80)	48(.83)	45(.81)
		1	1	142(.80)	140(.80)	138(.79)	136(.79)	148(.82)	142(.80)
			5	93(.81)	91(.81)	88(.79)	87(.79)	96(.82)	90(.80)
			10	87(.81)	85(.81)	81(.80)	81(.80)	89(.84)	83(.80)
	2	.1	1	22(.83)	21(.81)	22(.82)	20(.80)	23(.84)	23(.81)
			5	7(.83)	7(.83)	6(.78)	6(78)	7(.84)	7(.81)
			10	5(.82)	5(.83)	5(.83)	4(.77)	5(.83)	5(.82)
		0.5	1	34(.82)	33(.81)	32(.81)	31(.80)	34(.83)	34(.81)
			5	19(.84)	18(.82)	17(.82)	16(.80)	18(.83)	18(.82)
			10	17(.84)	16(.82)	15(.81)	15(.81)	16(.83)	16(.82)
		1	1	49(.83)	47(.82)	45(.81)	44(.80)	48(.83)	47(.81)
			5	35(.84)	33(.83)	30(.80)	29(.79)	32(.83)	31(.81)
			10	33(.86)	31(.84)	28(.81)	28(.81)	30(.83)	29(.82)
	3	0.1	1	8(.85)	8(.85)	8(.85)	7(.82)	8(.85)	8(.81)
			5	3(.88)	3(.87)	2(.76)	2(.78)	3(.89)	3(.87)
			10	2(.84)	2(.86)	2(87)	2(.87)	2(.87)	2(.84)
		0.5	1	14(.86)	13(.84)	12(.82)	11(.80)	12(.82)	13(.82)
			5	9(.89)	8(.86)	7(.83)	6(.78)	7(.83)	7(.81)
			10	8(.87)	7(.85)	6(.86)	6(.82)	7(.85)	7(.85)
		1	1	22(.87)	20(.85)	17(.81)	16(.80)	18(.83)	18(.81)
			5	17(.89)	15(.89)	12(.81)	11(.77)	13(.83)	13(.82)
			10	16(.89)	14(.87)	11(.80)	11(.80)	12(.82)	12(.81)
	4	0.1	1	5(.88)	4(.81)	5(.87)	4(.82)	4(.81)	5(.84)
			5	2(.90)	2(.91)	1(.68)	1(.73)	2(.92)	2(.90)
			10	2(.96)	1(.78)	1(.80)	1(.81)	2(.97)	2(.96)
		0.5	1	9(.87)	8(.86)	7(.80)	6(.79)	7(.82)	8(.83)
			5	7(.93)	5(.85)	5(.87)	4(.80)	5(.87)	5(.85)
			10	6(.92)	5(.88)	4(.84)	4(.84)	4(.83)	4(.82)
		1	1	15(.90)	13(.87)	10(.79)	9(.77)	11(.82)	12(.83)
			5	12(.91)	10(.88)	7(.79)	7(.80)	8(.83)	8(.82)
			10	12(.93)	10(.90)	7(.81)	7(.81)	8(.84)	8(.84)
1.2	1.5	0.1	1	64(.80)	65(.81)	66(.81)	63(.80)	67(.82)	68(.81)
			5	19(.82)	19(.81)	19(.81)	19(.82)	21(.85)	20(.81)
			10	14(.83)	14(.83)	13(.81)	13(.81)	15(.86)	14(.82)
		0.5	1	96(.81)	96(.81)	96(.81)	93(.80)	101(.82)	99(.80)
			5	51(.81)	51(.81)	50(.80)	49(.81)	53(.84)	51(.81)
			10	46(.82)	45(.81)	43(.80)	43(.80)	47(.83)	44(.80)
		1	1	136(.80)	136(.80)	134(.79)	131(.79)	144(.89)	137(.79)
			5	92(.82)	90(.81)	87(.80)	86(.80)	94(.81)	88(.80)
			10	86(.81)	84(.81)	81(.80)	81(.80)	87(.83)	82(.80)
	2	0.1	1	20(.82)	20(.82)	21(.83)	19(.81)	21(.84)	22(.82)
			5	6(.79)	6(.78)	6(.79)	6(.80)	7(.85)	7(.83)
			10	5(.84)	5(.84)	4(.78)	4(.79)	5(.84)	5(.83)
		0.5	1	32(.83)	32(.82)	31(.81)	29(.80)	32(.83)	32(.81)
			5	19(.84)	18(.83)	17(.82)	16(.79)	23(.90)	17(.81)
			10	17(.84)	16(.83)	15(.81)	15(.81)	16(.83)	16(.83)
		1	1	47(.82)	46(.82)	44(.81)	42(.80)	47(.83)	46(.80)
			5	34(.84)	32(.82)	29(.79)	29(.79)	32(.83)	31(.82)
			10	33(.85)	31(.84)	28(.81)	27(.79)	30(.83)	29(.82)

	3	0.1	1	7(.83)	7(.81)	8(.86)	6(.79)	7(.83)	8(.83)
			5	3(.90)	3(.90)	2(.78)	2(.80)	3(.91)	3(.88)
			10	2(.86)	2(.87)	2(.87)	2(.88)	2(.87)	2(.86)
		0.5	1	13(.85)	13(.86)	12(.82)	10(.78)	12(.84)	12(.80)
			5	9(.89)	8(.86)	7(.82)	6(.79)	9(.90)	7(.81)
			10	8(.87)	7(.84)	6(.81)	6(.82)	7(.85)	7(.84)
		1	1	21(.87)	19(.84)	17(.82)	16(.81)	18(.84)	18(.81)
			5	17(.90)	15(.87)	12(.80)	11(.78)	13(.83)	13(.82)
			10	16(.90)	14(.87)	11(.80)	11(.80)	12(.82)	12(.81)
	4	0.1	1	4(.83)	4(.83)	5(.89)	3(.77)	4(.83)	5(.87)
			5	2(.91)	2(.92)	1(.70)	1(.75)	2(.93)	2(.91)
			10	1(.76)	1(.79)	1(.80)	1(.82)	2(.97)	2(.97)
		0.5	1	9(.88)	8(.86)	7(.81)	6(.80)	7(.82)	8(.84)
			5	6(.92)	5(.86)	4(.80)	4(.81)	6(.92)	5(.86)
			10	6(.93)	5(.88)	4(.84)	4(.84)	4(.84)	4(.82)
		1	1	15(.90)	13(.87)	10(.80)	9(.78)	11(.83)	11(.80)
			5	12(.91)	10(.88)	7(.79)	7(.80)	8(.83)	8(.82)
			10	12(.93)	10(.90)	7(.81)	7(.81)	8(.85)	8(.84)

n_{w1}, n_{w2}, n_{lw1} and n_{lw2} are our methods and n_{exact} is the public method.

Sample size calculation based on testing multiple genes via FDR-controlling method

In this study, we evaluated the performance of the sample size methods based on testing the multiple genes via FDR-controlling method rather than the type I error $\boldsymbol{\alpha}$, which is widely used in the RNA-seq analysis. We set $\mathrm{m} = 10{,}000$, $m_1 = 100$, $m_0 = \mathrm{m} - m_1$ and want to detect the expected number of true DEG $t_1 = 80$ and the actual power corresponding to the nominal power of $1 - \beta = 80\%$. We also set $u_{0g} = 1, 5$ or 10, $\rho_g = 1.5, 2$, 3 or 4 and $\phi_g = 0.1, 0.5$ or 1.0 . With these settings, the new $\alpha^* = 4.25 \times 10^{-4}$ was obtained from the equation (14) at a desired FDR ($f = 0.05$). Then, we calculated the sample size by substituting α^* and power into the equations (15-18) and the published method using the exact test. For each designed setting, we also conducted 5000 simulations from an independent negative binomial distribution. The number of true DEGs was counted using p-values $\leq \alpha^*$ which is much smaller than the nominal type I error rate 0.05. The empirical power was obtained as the proportion of simulation samples for which H_0 is rejected with the nominal type I error $\alpha^* = 4.25 \times 10^{-4}$. The expected number of true DEGs under α^* can be estimated via the multiplication of the estimated power with the total number of true DEGs. The results in Table 3 report the estimated sample size with the empirical power given in parentheses under the cases $w = 1$ and 1.2.

Table 3: Simulated power corresponding to the samples size calculated for testing multiple genes. The sample sizes are calculated using six methods given a predefined $FDR = 0.05$ and 80% power.

w	ρ	ϕ	u_0	n_{w1}	n_{w2}	n_{lw1}	n_{lw2}	n_{lrt}	n_{exact}
1	*1.5*	*.1*	*1*	*197(.80)*	*196(.81)*	*198(.81)*	*192(.80)*	*204(.83)*	*200(.80)*
			5	58(.81)	57(.81)	57(.81)	55(.79)	60(.84)	58(.81)
			10	40(.81)	39(.80)	39(.80)	38(.79)	41(.84)	40(.81)
		0.5	1	288(.81)	284(.82)	283(.81)	277(.80)	293(.83)	286(.81)
			5	148(.82)	145(.81)	142(.80)	140(.80)	149(.83)	143(.81)
			10	131(.82)	127(.80)	124(.81)	123(.80)	131(.83)	125(.80)
		1	1	401(.81)	394(.81)	389(.81)	384(.80)	405(.82)	393(.80)
			5	262(.82)	255(.81)	248(.80)	247(.80)	261(.83)	250(.80)
			10	244(.83)	237(.82)	230(.81)	229(.81)	243(.83)	232(.82)
	2	.1	1	61(.82)	60(.81)	62(.83)	57(.80)	61(.83)	61(.81)
			5	19(.83)	18(.82)	18(.82)	17(.80)	19(.85)	19(.84)
			10	14(.84)	13(.82)	13(.82)	12(.80)	14(.86)	13(.82)
		0.5	1	96(.83)	92(.82)	91(.81)	86(.80)	93(.83)	91(.81)
			5	54(.85)	51(.84)	47(.81)	46(.80)	50(.83)	48(.81)
			10	49(.84)	45(.82)	42(.80)	41(.79)	44(.82)	43(.81)
		1	1	140(.83)	133(.81)	127(.80)	122(.80)	131(.82)	128(.79)
			5	98(.85)	91(.84)	84(.80)	83(.80)	88(.83)	85(.80)
			10	92(.84)	85(.83)	78(.81)	78(.81)	83(.83)	80(.81)
	3	0.1	1	22(.84)	21(.83)	22(.84)	18(.80)	21(.84)	21(.81)
			5	8(.88)	7(.83)	7(.84)	6(.79)	7(.85)	7(.83)
			10	6(.87)	5(.80)	5(.82)	4(.74)	5(.83)	5(.81)
		0.5	1	39(.86)	36(.85)	34(.83)	30(.80)	34(.84)	34(.82)
			5	25(.89)	22(.87)	18(.80)	18(.82)	20(.85)	19(.82)
			10	24(.91)	20(.87)	16(.79)	16(.80)	18(.85)	17(.80)
		1	1	61(.87)	54(.85)	48(.82)	44(.79)	49(.83)	49(.82)
			5	47(.91)	40(.87)	33(.81)	32(.80)	35(.83)	34(.81)
			10	45(.91)	38(.88)	31(.81)	30(.80)	33(.84)	32(.81)
	4	0.1	1	13(.85)	12(.83)	13(.85)	10(.81)	12(.86)	12(.82)
			5	5(.88)	5(.92)	4(.82)	3(.72)	4(.83)	4(.81)
			10	4(.88)	4(.93)	3(.83)	3(.87)	3(.83)	3(.81)
		0.5	1	26(.89)	23(.88)	20(.82)	17(.79)	20(.84)	20(.81)
			5	18(.93)	15(.91)	11(.79)	11(.81)	12(.83)	12(.81)
			10	17(.94)	14(.91)	10(.79)	10(.81)	11(.84)	11(.83)
		1	1	43(.93)	36(.89)	30(.82)	26(.79)	30(.83)	30(.81)
			5	35(.95)	28(.91)	20(.79)	20(.80)	22(.83)	22(.82)
			10	34(.95)	27(.92)	19(.80)	19(.81)	21(.84)	21(.83)
1.2	1.5	0.1	1	180(.80)	184(.80)	186(.81)	177(.80)	187(.82)	186(.81)
			5	54(.80)	55(.81)	54(.80)	52(.79)	56(.83)	55(.81)
			10	38(.80)	38(.80)	38(.81)	37(.80)	40(.84)	38(.80)
		0.5	1	270(.80)	273(.80)	271(.80)	262(.80)	280(.83)	272(.80)
			5	145(.82)	143(.81)	139(.81)	138(.81)	146(.84)	140(.81)
			10	129(.82)	126(.82)	123(.82)	122(.81)	128(.84)	122(.81)

		1	1	384(.81)	383(.81)	377(.81)	369(.80)	394(.83)	379(.80)
			5	258(.82)	253(.81)	245(.80)	244(.80)	256(.83)	247(.80)
			10	243(.83)	236(.82)	229(.80)	228(.80)	238(.84)	230(.81)
	2	0.1	1	55(.80)	57(.81)	59(.83)	52(.80)	57(.83)	57(.80)
			5	18(.83)	18(.83)	18(.83)	16(.80)	18(.84)	18(.83)
			10	13(.83)	13(.83)	12(.79)	12(.81)	13(.85)	13(.84)
		0.5	1	90(.82)	90(.82)	88(.81)	82(.81)	89(.83)	87(.80)
			5	53(.85)	50(.83)	47(.81)	45(.80)	49(.83)	47(.80)
			10	48(.84)	45(.83)	41(.79)	41(.80)	44(.83)	46(.85)
		1	1	134(.82)	130(.81)	124(.80)	118(.80)	128(.82)	124(.80)
			5	97(.84)	90(.83)	83(.80)	82(.79)	87(.83)	84(.80)
			10	92(.85)	85(.84)	78(.81)	77(.80)	82(.83)	79(.81)
	3	0.1	1	20(.85)	20(.83)	21(.85)	17(.81)	20(.86)	20(.83)
			5	7(.82)	7(.84)	7(.86)	6(.82)	7(.87)	7(.85)
			10	6(.89)	5(.81)	5(.84)	4(.76)	5(.85)	5(.83)
		0.5	1	37(.86)	35(84)	33(.82)	29(.81)	32(.83)	32(.80)
			5	25(.90)	22(.88)	18(.80)	17(.79)	19(.84)	19(.83)
			10	23(.90)	20(.88)	16(.79)	16(.80)	18(.85)	19(.87)
		1	1	59(.89)	53(.85)	47(.81)	43(.79)	49(.84)	48(.82)
			5	47(.91)	40(.88)	33(.82)	32(.81)	35(.84)	34(.82)
			10	45(.91)	38(.88)	31(.81)	30(.80)	33(.85)	32(.82)
	4	0.1	1	12(.87)	12(.86)	13(.88)	9(.90)	11(.85)	11(.80)
			5	5(.91)	4(.81)	4(.84)	3(.75)	4(.85)	4(.83)
			10	4(.90)	4(.94)	3(.83)	3(.87)	3(.84)	3(.82)
		0.5	1	25(.90)	22(.87)	20(.84)	16(.79)	19(.83)	20(.83)
			5	18(.94)	15(.91)	11(.80)	11(.82)	12(.84)	12(.83)
			10	17(.94)	14(.91)	10(.79)	10(.81)	11(.84)	12(.87)
		1	1	41(.92)	35(.88)	29(.81)	26(.81)	32(.88)	29(.80)
			5	35(.95)	28(.91)	20(.80)	20(.81)	22(.84)	22(.83)
			10	34(.95)	27(.92)	19(.80)	19(.80)	21(.84)	21(.83)

n_{w1}, n_{w2}, n_{lw1} and n_{lw2} are our methods and n_{exact} is the public method.

Given the sample size obtained from n_{lw2}, an empirical power for other methods is also computed from 5,000 simulations (Table 4). In order to compare the performance of these methods the paired Wilcoxon signed-rank test is used to test the simulated power in Table 3 for the statistical significance. The results are in Table 5.

Table 4: Simulated power for testing multiple genes given the sample size n_{lw2}, FDR = 0.05 and 80% power.

	ρ	ϕ	u_0	n_{lw2}	n_{lw1}	n_{w1}	n_{w2}	n_{lrt}	n_{exact}
w=1	1.5	.1	1	192(.80)	.79	.79	.79	.80	.78
			5	55(.79)	.78	.78	.78	.79	.78
			10	38(.79)	.78	.77	.78	.79	.78
		0.5	1	277(.80)	.79	.79	.78	.80	.78
			5	140(.80)	.79	.78	.78	.79	.79
			10	123(.80)	.80	.78	.80	.80	.80
		1	1	384(.80)	.80	.78	.78	.80	.79
			5	247(.80)	.80	.78	.79	.80	.80
			10	229(.81)	.81	.79	.80	.81	.80
	2	.1	1	57(.80)	.77	.77	.78	.78	.76
			5	17(.80)	.78	.76	.78	.79	.77
			10	12(.80)	.78	.74	.77	.78	.77
		0.5	1	86(.80)	.78	.76	.77	.78	.77
			5	46(.80)	.79	.74	.77	.79	.78
			10	41(.79)	.78	.72	.76	.78	.77
		1	1	122(.80)	.78	.74	.77	.78	.78
			5	83(.80)	.79	.73	.77	.79	.79
			10	78(.81)	.81	.74	.78	.80	.80
	3	0.1	1	18(.80)	.70	.71	.71	.76	.72
			5	6(.79)	.74	.66	.73	.76	.73
			10	4(.74)	.70	.53	.65	.71	.68
		0.5	1	30(.80)	.74	.66	.73	.76	.74
			5	18(.82)	.80	.63	.75	.79	.79
			10	16(.80)	.79	.59	.73	.78	.77
		1	1	44(.79)	.76	.62	.72	.76	.74
			5	32(.80)	.79	.57	.73	.78	.77
			10	30(.80)	.79	.55	.72	.78	.77
	4	0.1	1	10(.81)	.65	.62	.71	.75	.69
			5	3(.72)	.62	.39	.58	.67	.62
			10	3(.87)	.81	.62	.78	.83	.81
		0.5	1	17(.79)	.71	.52	.67	.74	.70
			5	11(.81)	.79	.41	.69	.78	.76
			10	10(.81)	.79	.33	.68	.77	.76
		1	1	26(.79)	.74	.44	.66	.74	.71
			5	20(.80)	.79	.34	.68	.77	.76
			10	19(.81)	.80	.30	.68	.79	.78

n_{w1}, n_{w2}, n_{lw1} and n_{lw2} are our methods and n_{exact} is the public method.

Table 5. P-values are calculated using the paired Wilcoxon signed-rank statistical test of the power values.

	$\boldsymbol{n_{lw1}}$	$\boldsymbol{n_{w2}}$	$\boldsymbol{n_{exact}}$	$\boldsymbol{n_{lrt}}$	$\boldsymbol{n_{w1}}$
$\boldsymbol{n_{lw2}}$	< .0001***	< .0001***	<.0001***	< .0001***	<.0001***
$\boldsymbol{n_{lw1}}$	-	< .001**	< .05***	< .0001***	< .0001***
$\boldsymbol{n_{w2}}$		-	< .001**	< .001*	< .05*
$\boldsymbol{n_{exact}}$			-	< .05*	< .001*
$\boldsymbol{n_{lrt}}$				-	<.0001***

*Statistically significant.

Identify the pattern of the sample sizes changing with different values of u_0, ϕ and ρ for different methods at α and α^* levels

First, we identified patterns of sample size changes with different values of u_0, ϕ and ρ for different methods. Although Tables 2-3 show similar pattern of sample sizes obtained from the six methods, we found that the required sample sizes $n_{lw1}$and n_{lw2} derived from the log-transformed Wald tests are the smallest compared with the other methods.

Figure 1 illustrates n_{lw2} varying with the values of u_0 when other parameters, such as ρ ∈(1.5, 2, 3, 4), ϕ ∈(.1, .5, 1) and w ∈(1, 1.2) are fixed. Given the nominal power (1-$\boldsymbol{\beta}$ = 0.8) and $\boldsymbol{\alpha}$ = 0.05 or FDR = 0.05, as expected, n_{lw2} decreases as u_0 increases under a fixed ρ and ϕ. This indicates that for a lowly expressed gene, a larger sample size is required to achieve a detection power of DEGs between two conditions. For a fixed u_0 and ρ, n_{lw2} increases as ϕ increases (Figure 1). This is also expected because a larger ϕ indicates higher variation of the genes across conditions. Furthermore, for a fixed ϕ and u_0, n_{lw2} decreases as ρ increases. This result indicates a smaller n is required for a larger difference of mean read counts between two conditions or vice versa. For the same setting of parameters, we found the n_{lw2} in Figure 1A and 1C with an equal size factor ($w = 1$) across conditions is slightly larger than the unequal size factor in Figure 1B ($w = 1$). Under the same settings, as expected, the sample sizes in Table 23with FDR = 0.05 are larger than those with $\boldsymbol{\alpha}$ = 0.05 (Table 2), indicating that a larger n is required for detecting DEGs while testing thousands of genes simultaneously (Figure 1C and 1D) compared with testing a single gene (Figure1A and 1B).

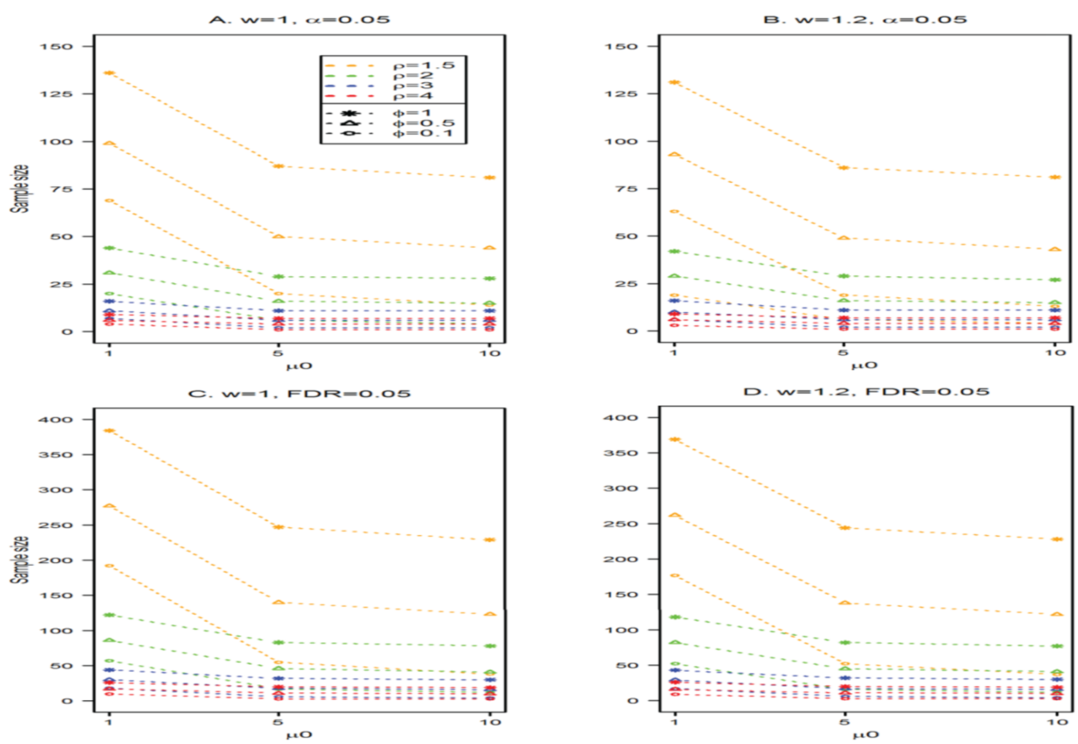

Figure 1: Sample size (n) for testing a single gene and multiple genes using the n_{lw2} method. n varies with mean reads counts in u_0 given the combination of ρ, ϕ, and w = 1 or 1.2 at 80% power and a nominal $\boldsymbol{\alpha}$ (0.05) for testing a single gene (Figures 1A and 1B)

or a nominal FDR (0.05) for testing multiple genes (Figures 1C and 1D). The fold changes (ρ = 1.5, 2, 3 and 4) are in orange, green, blue and red, respectively.

Comparison of the sample calculations from different methods based on testing a single genes in Table 2 and multiple genes in Table 3.

Next we compared the sample sizes (n) estimated from our five derived methods ($n_{w1}, n_{w2}, n_{lw1}, n_{lw2}, n_{lrt}$) and the public method (n_{exact}) [19]. Figure 2 illustrates that n decreases as u_0 increases for all methods given w = 1, power = 0.8, FDR = 0.05, $\phi \in (0.1, 0.5, 1)$ and $\rho \in (1.5, 2)$. Given $\rho = 1.5$, the sample sizes from all methods are getting close to each other when ϕ is 0.1 for small biological variation (Figures 2A-C). As ϕ increases, the difference between the sample sizes for all methods becomes much larger. Similar patterns were observed given $\rho = 2$ (Figures 2D-F). With regards to the n, we noticed that the sample sizes calculated from n_{lw1} and n_{lw2} methods (red in Figure 2) are the smallest while maintaining the nominal power close to or above 80%. Among the other four methods, n_{lrt} and n_{exact} performed better than n_{lw1} and n_{lw2} in most scenarios.

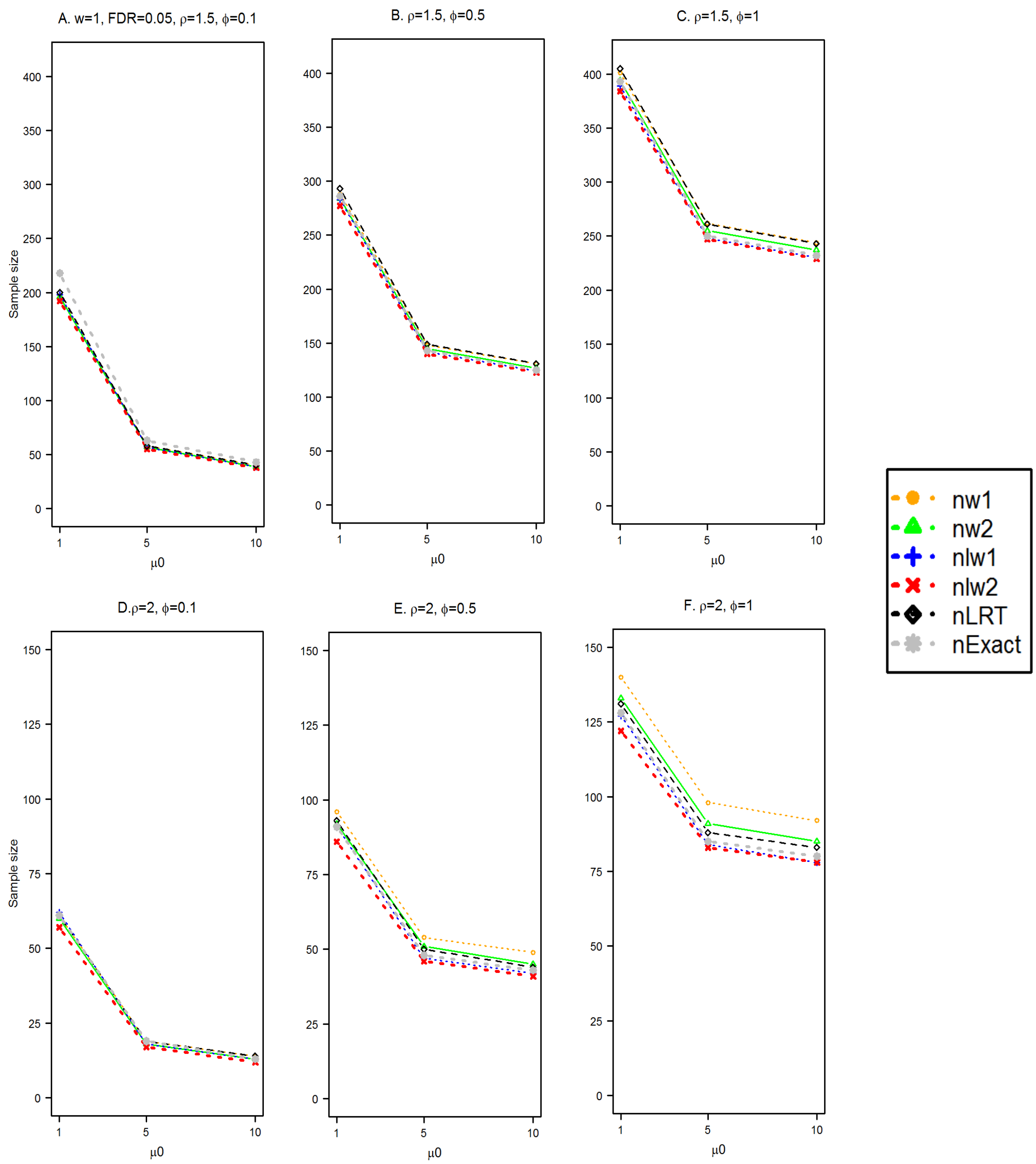

Figure 2: Sample size (n) for testing multiple genes using the six methods (n_{w1}, n_{w2}, n_{lw1}, n_{wl2}, n_{lrt} and n_{exact}). n varies with u_0 given the combination of ρ and ϕ at $w = 1$, 80% power and 0.05 FDR. n_{lw1} and n_{lw2} in red and blue respectively require smaller n.

Figures 2A and 2D illustrate samples sizes for all methods are close to each other when ϕ is 0.1.

In addition, the empirical power in parentheses (Tables 2 -3) was calculated from simulations with the size of 5,000 corresponding to the estimated n for all methods. The results show almost all of the methods are close to or higher than the desired power. Although the sample sizes calculated using n_{lw1} and n_{lw2} are the smallest, we cannot arbitrarily conclude that these two methods are the best because the corresponding empirical powers are varied corresponding to the sample sizes for each method (Tables 2-3).

For a better comparison with the same settings and a fixed and small n estimated from the log-transformed Wald method (n_{lw2}), we observed that n_{lw1} and n_{lw2} consistently achieve a better power close to the nominal power 80% or higher in all scenarios compared to other methods (Table 4). We also observed that n_{lrt} and n_{exact} perform similarly and both of them achieve a higher power than n_{w1} and n_{w2} when the fold change is great than 2. Table 5 from a paired Wilcoxon ranked test indicates that the empirical power from n_{lw2} is statistically significant from that achieved using other methods. Table 5 also shows that all the methods are significantly different from each other. Among these four methods, n_{lrt} and n_{exact} performed better than n_{lw1} and n_{lw2} with the power closer to an 80% nominal power in all scenarios (Table 4).

4. Application

Sample size calculation based on RNA-seq data in Human Breast Cancer: To identify DEGs between two conditions, we explored a real human breast cancer dataset to calculate the sample size. Forty Estrogen receptor positive (ER+) and HER2 negative breast cancer primary tumors and 29 uninvolved breast tissue sample that were adjacent to ER+ primary tumors in .fastq format were downloaded from NCBI GEO (series ID GSE58135). The raw sequencing files were mapped to the human hg19 reference genome using tophat2 (v2.0.13) with bowtie version (2.2.3.0). The mapped counts for each gene per sample were then extracted using HTSeq-scripts-count (version 2.7). There are a total of 57,773 genes extracted. After filtering genes with the mean read counts less than one in two groups, 35,112 genes were left. These samples were loaded into *edgeR* to estimate common dispersion and size factors. With the aid of *edgeR*, the normalization factors called size factors are estimated using the "RLE" scaling factor method [17] and the estimated ratio of total size factors of the samples in each condition (w) is 1.1. The common dispersion ϕ is estimated as 0.48 using the CMLE method [22].

We assumed the top 400 of 35,122 genes (1.1%) are prognostic and have the largest fold changes. The minimum of average read counts among these genes in the control group served as pilot data was estimated as $u_0 = 1$ [16]. In addition, the sample sizes were estimated using $u_0 = 5, 10$ and 20. Suppose we want to set the nominal power to 80%, which indicates we want to identify 320 or more of the prognostic genes. Under the control FDR at $f = 0.10$ and 80% power, we can set $\mathrm{m} = 35{,}112$, $m_1 = 400$, $m_0 = \mathrm{m} - m_1$ and $t_1 = 320$. The parameters ρ_g $(g = 1, \dots, 35{,}112)$ were assumed to be unknown. Let the mean counts in the control condition $u_{0g} = 1, 5, 10$, or 20 and fold change $\rho_g = 1.5$ or 2, 3 and 4 with common dispersion $\phi_g = 0.48$. With these settings, the new $\alpha^* = 9.992 \times 10^{-4}$ was obtained from equation (24) at a desired FDR($f =$

0.10). Then, we calculated the sample size by substituting α^* and the power into equations (25-29) for each method (Table 6).

Table 6 reports the samples sizes under different scenarios including various minimum mean read counts in control condition (1, 5, 10 and 20) and desired fold changes (1.5, 2, 3 and 4) while controlling the FDR at 0.10. We found that the original RNA-seq experiment [28] with a minimum sample size 31 in each condition can detect more than 80% of the prognostic genes at the FDR (f = 0.10) and $u_0 = 1$ if the desired fold change is 3 or more. Moreover, with a minimum sample size 42 in each condition, we found that it can detect more than 80% of the prognostic genes at the FDR (f = 0.10) and $u_0 = 5$ if the desired fold change is 2 or more.

Table 6: The sample sizes per group required for a balanced design in human breast cancer RNA-seq data. The sample sizes are calculated given the nominal $FDR = 0.10$ and 80% power.

	ρ	u_0	n_{w1}	n_{w2}	n_{lw1}	n_{lw2}	n_{lrt}	n_{exact}
w =1.1 ϕ=.48	1.5	1	242	241	240	234	250	242
		5	125	123	120	119	126	121
		10	111	108	105	104	111	106
		20	103	101	97	97	103	99
	2	1	80	79	77	73	78	78
		5	46	43	40	39	42	41
		10	41	39	35	35	38	36
		20	39	36	33	33	35	34
	3	1	33	31	29	25	29	29
		5	21	19	16	15	17	16
		10	20	17	14	14	15	15
		20	19	16	13	13	**14**	14
	4	1	22	19	18	15	17	17
		5	16	13	10	9	11	10
		10	15	12	9	8	10	10
		20	14	11	8	8	9	9

* The dispersion $\phi = 0.48$ and the ratio of the size factor $w = 1.1$ are estimated using *edgeR* package particularly for RNA-seq data.

5. Discussion and conclusion

In this study, five methods ($n_{w1}, n_{w2}, n_{lw1}, n_{lw2}$ and n_{lrt}) were derived to calculate sample sizes using the Wald test and LRT statistics based on a negative binomial distribution for modeling an RNA-seq experiment. The parameters are estimated using the MLE and CMLE methods. Since the dispersion estimated from MLE has no closing form, it is difficult to derive the sample size formula. Therefore, all of the methods are based on a fixed and constant dispersion. A log-transformed approach corresponding to the modified $Z_{w1}$and Z_{w2} was used to derive two other test statistics ($Z_{lw1}$and Z_{lw2}). For all these statistical tests as well as the exact test [19] that are used to derive the sample size calculation formulas, gene expression levels are assumed to be independent in each sample. Although this assumption might not hold in reality, it is widely used in RNA-seq as well as in microRNA data analysis. In this study, we assume equal sample size in the two conditions to derive the sample size formula. The derived formula for sample size calculations can be easily extended to the unequal sample sizes by setting n_1 = kn_0. In our simulation study, we set the ratio of total size factors in two conditions as 1 and 1.2 instead of $w = 2$ in the study [19]. In reality, the read depths of RNA-seq samples generated from the same run are very close to each other across conditions. Therefore, we think $w = 1.2$ or close to 1 is more common than $w = 2$. Furthermore, in our simulation and application studies, the minimum sample sizes required to achieve a nominal power of 80% with a predefined FDR (f = 0.05 or 0.10) are usually larger than those in an RNA-seq experiment due to the real costs. In such a situation, we can increase the read depth per sample to indirectly increase the mean of read counts u_0 in the control condition. Thus, the required sample sizes can be decreased correspondingly.

Among the methods we evaluated, the simulation results show that n_{lw2} from the log transformed Wald test with the parameters estimated from CMLE is the best method because a smaller sample size is required for designing an RNA-seq experiment while achieving a power close to or higher than 80% at a pre-defined FDR = 0.05. The second best method is n_{lw1} with the parameters estimated from unrestricted MLEs. We also found that n_{lrt} and n_{exact} methods perform better than n_{w1} method based on the estimated power given the genes with a fold change > 2. However, n_{w1} achieve a better power than n_{lrt} and n_{exact} given a fold change ≤ 2. In summary, n_{w1}, n_{w2}, n_{lrt} and n_{exact} methods varied with different scenarios. Finally, since the log-transformed sample size calculation methods are more robust, simpler and require less time, we hope our tables can help and benefit for researchers and scientists in the design of RNA-seq experiments.

SAMPLE SIZE CALCUALTION METHODS FOR RNA-SEQ DATA USING A GENERALIZED LINEAR MODEL

1. Introduction

High-dimensional RNA-seq is a powerful tool for gene expression profiling and has been increasingly used for biomarker discovery and clinical trial studies [29, 30]. Differentially expressed genes (DEGs) or novel transcripts identified by RNA-seq have served as gene signatures for clinical diagnosis, prognosis, and to estimate the efficacy of gene therapy or survival prediction [31-35]. With the rapid growth of clinical trial applications using RNA-seq, sample size estimation methods are critically needed for experimental design.

Since RNA-seq data are read counts, sample size calculations were proposed and derived on the basis of a Poisson distribution using a Wald's test, a score test and a likelihood ratio test (LRT) statistic in the case of testing a single gene, respectively [13, 16]. In fact, a sample size formula derived from a score statistic for a Poisson distribution is equivalent to the method that was derived from the Wald's test statistic with parameters estimated from a constrained maximum likelihood estimate (MLE) under the same null hypothesis (H_0). In addition, Li *et al.* [16] introduced a false discovery rate (FDR) when the sample size formula was derived from a score test statistics with a Poisson distribution in the case of testing multiple genes. However, studies have shown that the variance of the read counts of genes from RNA-seq data with biological replicates is significantly larger than the mean, resulting in an over-dispersed data [17, 36]. Therefore, the variance needs to be adjusted for proper inference. Consequently, a negative binomial (NB) distribution with two-parameters has been proposed and used for modeling RNA-seq count data. However, sample size calculations from a NB distribution for RNA-seq data is more challenging when compared with that using a Poisson distribution. Recently, Li *et al.* proposed another sample size calculation method using a NB distribution and an exact test with the aid of *edgeR* in the case of testing a single gene and multiple genes between two groups [19]. Most recently, several sample size methods based on a NB distribution using a Wald's test, a log-transformed Wald's test and a likelihood ratio test statistics with a one-sided test p-value were proposed by Li *et al.* [37]. However, all these methods assumed independent expression of genes between sample groups. In addition, the Wald's test statistic was based on the fold change (FC) or on the ratio of the true expression while the variance was estimated from the formula as $\sigma_T^2 = \frac{\gamma_1}{s_1} + \frac{\gamma_0}{s_0} + \frac{\phi}{n}(\gamma_1^2 + \gamma_0^2)$, in which the parameters γ_0 and γ_1 were estimated by an unrestricted MLE or constrained MLE (CMLE) under H_0 and the dispersion parameter ϕ was assumed to be a constant. In addition, all of these sample size methods assumed equal sample sizes for the two comparison groups. Although imbalanced designs are not popular for RNA-seq studies, in some situations, especially in the clinical studies, they are preferred. For example, in some cancer datasets, sample sizes in a control group are usually smaller than those in the cancer patient group.

Explicit sample size formulas analyzing rates of asthma or chronic obstructive pulmonary disease (COPD) in clinical studies has been proposed, which used a standard generalized linear model (GLM) with a NB distribution for equal or unequal sample size design [38, 39]. We use a similar approach to derive two explicit sample size calculation methods for RNA-seq data using a two-sided Wald's test that was based on the GLM. The variance in the Wald's test statistic used the variance-covariance matrix calculated from the parameters that were estimated from the MLE and the CMLE under two scenarios [39]. Using this approach, our method is able to incorporate the ratio of size factors for correcting sequencing read depth and the fold change for DEGs at a two-sided test statistic. Moreover, the sample size formula is applicable for both balanced and imbalanced designs. Since tens of thousands of genes in a RNA-seq sample are tested simultaneously for DEG analysis, a FDR is commonly used to adjust p-values in the multiple comparisons [16, 19]. Therefore, a much smaller type I error alpha (α^*) was calculated and incorporated in our sample size formula in the case of testing multiple genes.

In summary, this study is motivated to develop new sample size calculation formulas using a GLM for RNA-seq data that are applicable in different scenarios, such as testing a single gene or multiple genes, imbalanced or balanced design, equal or unequal read depth across samples, and a two-sided test p-values. A simulation study under different parameter settings is carried out to evaluate our formulas from a Wald's test versus an existing method that used an exact test. In section 4, we illustrate how to estimate sample sizes using publicly accessible RNA-seq data on breast cancer.

2. Sample size calculation methods based on a GLM

2.1 Sample size formulas in the case of testing a single gene

For a single gene in RNA-seq data, suppose that independent random sample Y_{ij} from the j-th sample $(j = 1, \dots, n_i)$ and in the i-th treatment group $(i = 0, 1)$ has an identical negative binomial distribution: $Y_{ij} \sim NB(s_{ij}\gamma_i, \phi)$ or $Y_{ij} \sim NB(u_{ij}, \phi)$[17, 22], where γ_i is the true gene expression level, s_{ij} is a size factor to normalize the raw reads in the different number of mapped reads, $u_{ij} = s_{ij}\gamma_i$ is the mean read count, and ϕ is a common dispersion parameter. Thus, the summation of reads per gene per group $(Y_i = \sum_{j=1}^{n_i} Y_{ij})$ also follows a negative binomial distribution with parameters $n_i u_i = t_i = n_i \bar{s}_i \gamma_i$ and ϕ/n_i, where $\bar{s}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} s_{ij}$ is the mean of the size factor for mapping reads in condition i and n_i is the number of biological replicates with the assumption $k = n_1/n_0$. The probability density function (PDF) of the observation y_{ij} for a NB distribution is [38, 39]:

$$\mathrm{P}(y_{ij}) = \frac{\Gamma(\phi^{-1} + y_{ij})}{\Gamma(\phi^{-1}) y_{ij}!} \left(\frac{\phi u_{ij}}{1 + \phi u_{ij}}\right)^{y_{ij}} \left(\frac{1}{1 + \phi u_{ij}}\right)^{\phi^{-1}}, \tag{20}$$

where the variance of y_{ij} is $\sigma_{ij}^2 = u_{ij} + \phi u_{ij}^2$. The above NB model utilizes the conventional parameterization called NB2 because we only know the mean read counts (u_{ij}) of the gene and don't know the true expression (γ_i) in the RNA-seq data [40].

For a GLM, the expected mean read counts (u_{ij}) of y_{ij} can be modeled by a log link function as:

$$\log u_{ij} = \log s_{ij} + \log \gamma_i = \log s_{ij} + \beta_0 + \beta_1 x_{ij},$$

where x_{ij} is the treatment group indicator, with x_{ij}= 0 if $i = 0$ for the control group, and x_{ij}= 1 if $i = 1$ for the treatment group; the quantity log s_{ij} denotes an offset. We are interested in the true mean per size factor unit:

$$\log \gamma_i = \beta_0 + \beta_1 x_i , \qquad (21)$$

where $\gamma_i = \frac{u_i}{\bar{s}_i}$, u_i denotes the expected mean read counts for gene g in group i, $\bar{s}_i$ denotes the mean of the size factors in group i. Thus, the true expression γ_0 and γ_1 from equation (21) can be obtained by:

$$\gamma_0 = e^{\beta_0}, \gamma_1 = e^{\beta_0+\beta_1} \text{ and } \frac{\gamma_1}{\gamma_0} = e^{\beta_1}.$$

For the detection of a DEG from RNA-seq data, the ratio $\rho = \gamma_1/\gamma_0$ typically represents the fold change. If it equals to one, this gene is not differentially expressed. Therefore, we are interested in testing:

$$H_0: \gamma_1 = \gamma_0 \text{ vs. } H_1: \gamma_1 \neq \gamma_0,$$

and making inference about the ratio ρ using the Wald's test statistic. This is equivalent to test:

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0. \qquad (22)$$

Replacing $u_{ij} = s_{ij}e^{\beta_0+\beta_1 x_{ij}}$ in equations (20), the log-likelihood function is:

$$l\,(\beta_0, \beta_1, \phi | y_{ij}, x_{ij}) = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \left[\log \frac{\Gamma(\phi^{-1}+y_{ij})}{\Gamma(\phi^{-1}) y_{ij}!} + y_{ij} \log\,(\phi s_{ij} e^{\beta_0+\beta_1 x_{ij}}) - \left(y_{ij} + \frac{1}{\phi}\right) \log\left(1 + \phi s_{ij} e^{\beta_0+\beta_1 x_i}\right) \right]. \tag{23}$$

The model parameters β_0, β_1 and ϕ in equation (23) can be estimated using the maximum likelihood estimates (MLE) and the variance-covariance matrix of model can be estimated via the inverse of the Fisher information matrix, asymptotically. Since the dispersion ϕ does not have a closed form, for simplicity, ϕ was treated as a constant for the two groups.

Deriving Fisher's information matrix: The first and second partial derivative of the log likelihood function (l) in equation 22) with respect to β_1 and β_2 are:

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} + 1/\phi) \left(\frac{d_{ij}}{1+d_{ij}}\right), \tag{24}$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij} - \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} + 1/\phi) \left(\frac{x_i d_{ij}}{1+d_{ij}}\right), \tag{25}$$

$$\frac{\partial^2 l}{\partial \beta_0^2} = -\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} + 1/\phi) \left[\frac{d_{ij}}{(1+d_{ij})^2}\right], \tag{26}$$

$$\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} + 1/\phi) \left[\frac{x_i d_{ij}}{(1+d_{ij})^2}\right], \tag{27}$$

$$\frac{\partial^2 l}{\partial \beta_1^2} = -\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} + 1/\phi) \left[\frac{{x_i}^2 d_{ij}}{(1+d_{ij})^2}\right], \tag{28}$$

where $d_{ij} = \phi s_{ij} e^{\beta_0+\beta_1 x_i}$ and assuming ϕ is constant.

The Fisher information matrix is defined as

$$I_{n_0,n_1}(\beta_0, \beta_1) = E\left[-\frac{\partial^2 l}{\partial(\beta_0,\beta_1)^2}\right] = \begin{bmatrix} \xi_{\beta_0\beta_0} & \xi_{\beta_0\beta_1} \\ \xi_{\beta_1\beta_0} & \xi_{\beta_1\beta_1} \end{bmatrix}.$$

From equations (24-28), we can obtain by simplification:

$$\xi_{\beta_0\beta_0} = -E\left[\frac{\partial^2 l}{\partial \beta_0^2}\right] = \frac{n_0\bar{s}_0 e^{\beta_0}}{1+\phi\bar{s}_0 e^{\beta_0}} + \frac{n_1\bar{s}_1 e^{\beta_0+\beta_1}}{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}},$$

$$\xi_{\beta_0\beta_1} = -E\left[\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1}\right] = \frac{n_1\bar{s}_1 e^{\beta_0+\beta_1}}{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}},$$

$$\xi_{\beta_1\beta_1} = -E\left[\frac{\partial^2 l}{\partial \beta_1^2}\right] = \frac{n_1\bar{s}_1 e^{\beta_0+\beta_1}}{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}},$$

where $E\left(\sum_{j=1}^{n_0} y_{0j}\right) = n_0\bar{s}_0 e^{\beta_0}$, $E\left(\sum_{j=1}^{n_1} y_{1j}\right) = n_1\bar{s}_1 e^{\beta_0+\beta_1}$. We also set $s_{0j} = \bar{s}_0$ and $s_{1j} = \bar{s}_1$ for simplification.

The Fisher information matrix can be expressed as:

$$I = E\left(-\frac{\partial^2 l}{\partial^2(\beta_0,\beta_1)}\right) = \begin{pmatrix} A+B & B \\ B & B \end{pmatrix}, \tag{29}$$

where $A = \frac{n_0\bar{s}_0 e^{\beta_0}}{1+\phi\bar{s}_0 e^{\beta_0}}$ and $B = \frac{n_1\bar{s}_1 e^{\beta_0+\beta_1}}{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}}$.

Let $(\hat{\beta}_0, \hat{\beta}_1)'$ be the MLEs of model parameters $(\beta_0, \beta_1)'$. The asymptotic normality and consistency of (β_0, β_1) follow the large-sample properties of MLE; and the variance-covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)'$ is the inverse of information matrix of $I_{n_0,n_1}(\beta_0, \beta_1)$ asymptotically and it is:

$$\Sigma(\hat{\beta}_0, \hat{\beta}_1) = I^{-} = \begin{bmatrix} 1/A & -1/A \\ -1/A & 1/A + 1/B \end{bmatrix}.$$

Thus, the variance of $\hat{\beta}_1$ is

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{1}{A} + \frac{1}{B} = \frac{1+\phi\bar{s}_0 e^{\beta_0}}{n_0\bar{s}_0 e^{\beta_0}} + \frac{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}}{n_1\bar{s}_1 e^{\beta_0+\beta_1}} = \frac{1}{n_0}\left(\frac{1+\phi\bar{s}_0 e^{\beta_0}}{\bar{s}_0 e^{\beta_0}} + \frac{1+\phi\bar{s}_1 e^{\beta_0+\beta_1}}{k\bar{s}_1 e^{\beta_0+\beta_1}}\right), \tag{30}$$

where $k = n_1/n_0$. Replacing $\gamma_0 = e^{\beta_0}, \gamma_1 = e^{\beta_0+\beta_1},\ \rho = \frac{\gamma_1}{\gamma_0} = e^{\beta_1}, w = \frac{\bar{s}_1}{\bar{s}_0}, u_0 =$ $\bar{s}_0\gamma_0$ and $u_1 = \bar{s}_1\gamma_1 = w\rho u_0$ in equation (30), the variance of $\hat{\beta}_1$ becomes

$$\text{Var}\left(\hat{\beta}_1\right) = \frac{1}{n_0}\left(\frac{1+\phi\bar{s}_0\gamma_0}{\bar{s}_0\gamma_0} + \frac{1+\phi\bar{s}_1\gamma_1}{k\bar{s}_1\gamma_1}\right) = \frac{1}{n_0}\left(\frac{1+\phi u_0}{u_0} + \frac{1+\phi u_1}{ku_1}\right) = \frac{1}{n_0}\left(\frac{1+\phi u_0}{u_0} + \frac{1+\phi w\rho u_0}{kw\rho u_0}\right)$$

$$= \frac{1}{n_0}\left[\frac{1}{u_0}\left(1 + \frac{1}{kw\rho}\right) + (1 + \frac{1}{k})\phi\right], \tag{31}$$

where *w* denotes the ratio of mean size factor for a gene g in condition *i* and ρ denotes a fold change.

Wald's Test statistic: To test the null hypothesis H_0: $\beta_1 = 0$, $\hat{\beta}_1$ asymptotically follows a normal distribution with mean $\log(\rho) = \log(\frac{\gamma_1}{\gamma_0})$ and variance Var $\left(\hat{\beta}_1\right)$ with $\rho = 1$ under H_0. The two-sided Wald's test is to reject H_0 if

$$|Z_w| = \frac{|\hat{\beta}_1|}{\sqrt{V_0(\hat{\beta}_1)}} > Z_{1-\frac{\alpha}{2}}. \tag{32}$$

Thus, a power of $(1 - \boldsymbol{\beta})$ is obtained under the alternative hypothesis $H_1 : \beta_1 \neq 0$. When $|Z_w| > Z_{1-\alpha/2}$, an approximately 2-sided Wald's test of $\hat{\beta}_1$with an expected mean $\log(\rho) = \log(\frac{\gamma_1}{\gamma_0})$and variance V_1 can be expressed as:

$$\frac{|\ \hat{\beta}_1 - E(\hat{\beta}_1)|}{\sqrt{V_1(\hat{\beta}_1)}} > \left[\frac{Z_{1-\alpha/2}\sqrt{V_0(\hat{\beta}_1)} - \log\left(\frac{\gamma_1}{\gamma_0}\right)}{\sqrt{V_1(\hat{\beta}_1)}}\right].$$

Hence, the approximate power can be expressed as:

$$1-\beta = 1-\phi\left[\frac{Z_{1-\alpha/2}\sqrt{V_0(\hat{\beta}_1)}-\sqrt{n_0}\log\left(\frac{\gamma_1}{\gamma_0}\right)}{\sqrt{V_1(\hat{\beta}_1)}}\right].$$

Solving the above equation , we obtain the sample size formula as:

$$n_0 = \frac{(Z_{1-\alpha/2}\sqrt{V_0}+Z_{1-\beta}\sqrt{V_1})^2}{\{\log(\rho)\}^2}, \tag{33}$$

where Z_p is the quantile of the standard normal distribution, $\rho = \frac{\gamma_1}{\gamma_0}$ denotes a fold change of a gene between two groups under H_1, V_0 and V_1 are the corresponding variance proportion estimated in equation (10) under H_0 and H_1, respectively, and defined as:

$$V_0 = \frac{1}{\tilde{u}_0}\left(1+\frac{1}{kw\rho}\right)+\left(1+\frac{1}{k}\right)\phi, \tag{34}$$

$$V_1 = \frac{1}{u_0}\left(1+\frac{1}{kw\rho}\right)+\left(1+\frac{1}{k}\right)\phi, \tag{35}$$

where $\tilde{u}_0$ and u_0 are the estimated and the true mean read counts under H_0 and under H_1, respectively.

Estimation of variance under H_0: The mean read counts $\tilde{u}_0$ in the control group in equation (34) can be estimated through the following approaches that are similar to the methods by Zhu *et al.* [39].

Method 1 (M1): Use the true mean read counts u_0

Let $\tilde{u}_0$ equal to the true u_0 and $\tilde{u}_1$ equal to the true $\rho w u_0$, which is equivalent to $\tilde{\gamma}_0 = \gamma_0$ and $\tilde{\gamma}_1 = \gamma_1 = \rho\gamma_0$. Then, we have $V_{0M1} = V_1$.

Thus the sample size formula under $V_{0M1} = V_1$ in equation (33) becomes:

$$n_{M1} = \frac{\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2 V_1}{\{\log(\rho)\}^2}, \tag{36}$$

where $V_1 = \frac{1}{u_0}\left(1 + \frac{1}{kw\rho}\right) + \left(1 + \frac{1}{k}\right)\phi = \frac{1}{\bar{y}_0}\left(1 + \frac{1}{kw\rho}\right) + \left(1 + \frac{1}{k}\right)\phi$.

These true parameters can be estimated by the unconstrained maximum likelihood estimation (MLE) similar to the method by Li *et al.*[37]. We obtain $\tilde{u}_0 = \bar{y}_0$ and $\tilde{u}_1 = \rho w \tilde{u}_0$, where $\tilde{u}_0$ is estimated from the sample mean ($\bar{y}_0$) in the control group [23, 37].

Method 2 (M2): Use the MLE under H_0

Under H_0: $\rho = \frac{\gamma_1}{\gamma_0} = 1$, we used a similar approach as that in Zhu *et al.* to estimate $\tilde{\gamma}_0$ and $\tilde{u}_0$.

Calculating MLEs of true expression parameter (γ_0 and γ_1) under unrestricted and restricted scenarios: The log likelihood function (l) with respect to γ_0 and γ_1 in equation (22) can be written as

$$l = \sum_{i=0}^{1} \sum_{j=1}^{n_i} \left[\log \frac{\Gamma(\phi^{-1}+y_{ij})}{\Gamma(\phi^{-1}) y_{ij}!} + y_{ij} \log \phi s_{ij}\gamma_i - \left(y_{ij} + \frac{1}{\phi}\right) \log\left(1 + \phi s_{ij}\gamma_i\right)\right].$$

Unrestricted scenario: The MLEs of $(\hat{\gamma}_0, \hat{\gamma}_1)'$ can be estimated by setting the partial derivative of *l* in B1 with the respect to $(\gamma_0, \gamma_1)'$ to zero and then solving.

$$\frac{\partial l}{\partial \gamma_0} = \frac{1}{\gamma_0} \sum_{j=1}^{n_0} y_{0j} - \frac{\phi \bar{s}_0}{1+\phi \bar{s}_0 \gamma_0} \sum_{j=1}^{n_0} y_{0j} - \frac{n_0 \bar{s}_0}{1+\phi \bar{s}_0 \gamma_0}, \tag{37}$$

$$\frac{\partial l}{\partial \gamma_1} = \frac{1}{\gamma_1} \sum_{j=1}^{n_1} y_{1j} - \frac{\phi \bar{s}_1}{1+\phi \bar{s}_1 \gamma_1} \sum_{j=1}^{n_1} y_{1j} - \frac{n_0 \bar{s}_1}{1+\phi \bar{s}_1 \gamma_1}, \tag{38}$$

where setting $s_{0j} = \bar{s}_0$and $s_{1j} = \bar{s}_1$ for simplification.

Setting equations 37 and 38 to zero and solving, we get $\hat{\gamma}_0 = \frac{\bar{y}_0}{\bar{\bar{s}}_0}$, $\hat{\gamma}_1 = \frac{\bar{y}_1}{\bar{\bar{s}}_1}$, $\hat{u}_0 = \bar{y}_0$ and $\hat{u}_1 = \bar{y}_1$.Thus, the MLEs of $(\hat{\beta}_0, \hat{\beta}_1)$ are $\hat{\beta}_0 = log(\hat{\gamma}_0) = log\left(\frac{\bar{y}_0}{\bar{\bar{s}}_0}\right)$, and $\hat{\beta}_1 = log\left(\frac{\hat{\gamma}_1}{\hat{\gamma}_0}\right) = log\left(\frac{\bar{y}_1}{w\bar{y}_0}\right)$.

Restriction scenario: Under null hypothesis (H_0): $\gamma_0 = \gamma_1$, the parameter γ_1 for *l* in (37) is replaced by γ_0. We take the partial derivative of *l* with respect to γ_0 and get

$$\frac{\partial l}{\partial \gamma_0} = \frac{1}{\gamma_0}\sum_{j=1}^{n_0} y_{0j} - \frac{\phi\bar{s}_0}{1+\phi\bar{s}_0\gamma_0}\sum_{j=1}^{n_0} y_{0j} - \frac{n_0\bar{s}_0}{1+\phi\bar{s}_0\gamma_0} + \frac{1}{\gamma_0}\sum_{j=1}^{n_1} y_{1j} - \frac{\phi\bar{s}_1}{1+\phi\bar{s}_1\gamma_0}\sum_{j=1}^{n_1} y_{1j} - \frac{n_0\bar{s}_1}{1+\phi\bar{s}_1\gamma_0}. \quad (39)$$

For simplification, we set $\bar{s}_0 = \bar{s}_1 = \bar{s}$, and $\bar{s} = \frac{\bar{s}_0+\bar{s}_1}{2} = \frac{\bar{s}_0(1+w)}{2}$. Thus, we get $\tilde{\gamma}_0$ and $\tilde{u}_0$ by setting equation 39 to zero and solving. Thus, under H_0, the MLEs of $\tilde{\gamma}_0$ and $\tilde{u}_0$ are

$$\tilde{\gamma}_0 = \frac{n_0\bar{y}_0+n_1\bar{y}_1}{(n_0+n_1)\bar{s}} = \frac{n_0\bar{s}_0\gamma_0+n_1\bar{s}_1\gamma_1}{(n_0+n_1)\frac{\bar{\bar{s}}_0(1+w)}{2}} = \frac{2(1+kw\rho)\gamma_0}{(1+k)(1+w)}, \quad (40)$$

$$\tilde{u}_0 = \bar{s}_0\tilde{\gamma}_0 = \frac{2(1+kw\rho)u_0}{(1+k)(1+w)}, \quad (41)$$

where u_0, the true mean read counts in the control group, can be estimated from the sample mean. Thus, we have $V_{0M2} = \frac{1}{\tilde{u}_0}\left(1+\frac{1}{kw\rho}\right) + \left(1+\frac{1}{k}\right)\phi$.Thus, the sample size formula under $V_0 = V_{0M2}$ in (33) becomes:

$$n_{M2} = \frac{\left(Z_{1-\alpha/2}\sqrt{V_{0M2}}+Z_{1-\beta}\sqrt{V_1}\right)^2}{\{\log(\rho)\}^2}. \quad (42)$$

2.2 Sample size formulas in the case of testing multiple genes simultaneously

The sample size formulas for a single gene are defined in equations (36) and (42), where $\boldsymbol{\alpha}$ is the significance level and $(1-\beta)$ is the power in percentile. In the case of testing multiple genes, replacing α with α^* in equations (36, 42), the corresponding sample size calculation formulas corrected by FDR at level f are, respectively,

$$n_{M1*} = \frac{\left(Z_{1-\alpha^*/2}+Z_{1-\beta}\right)^2 V_1}{\{\log(\rho)\}^2}, \tag{43}$$

$$n_{M2*} = \frac{\left(Z_{1-\alpha^*/2}\sqrt{V_{0M2}}+Z_{1-\beta}\sqrt{V_1}\right)^2}{\{\log(\rho)\}^2}. \tag{44}$$

3. Simulation studies for performance evaluation and results

Simulations were carried out to evaluate the performance of our sample size calculation methods and the published method. It includes two parts: in the first part, sample sizes and power are estimated in case of testing a single gene under a balanced design; in the second part, sample sizes and power are estimated in the case of testing multiple genes under a balanced or imbalanced experimental design with different read depth.

3.1. Sample size and power calculations in the case of testing a single gene

Simulation study: In this simulation, the mean read counts u_0 in the control group are set to be 1, 5, 10 or 50; the fold change $\rho=\frac{\gamma_1}{\gamma_0}$ is set to be

$0.5, 0.8, 1.2, 1.5, 2$ or 3; the ratio of mean size factors $w = \frac{\bar{s}_1}{\bar{s}_0}$ is set to be 1 for equal read depth; the constant dispersion parameter ϕ is set to be $0.1, 0.5$ or 1; the ratio of sample sizes is set to be $k = \frac{n_1}{n_0} = 1$ or $\frac{3}{2}$, where $k = 1$ is for the balanced design.

For each combination of these parameter settings, sample sizes in the case of testing a single gene were calculated using our formulas in (36) and (42), and the existing method for an 80% nominal power and a two-sided Wald's test at the significance level of α=0.05, respectively. The corresponding power for the estimated sample size was calculated via simulation using the Wald's test in (34).

Given the estimated sample size and each designed parameter setting, two sets of NB random variables were generated for two treatment groups: the control and the experimental group using an R script. For the control group, the random samples were generated given the parameters n_0, u_0 and ϕ. For the treatment group, the random samples are generated given the parameters $n_1 = kn_0$, $u_1 = \rho w u_0$ and ϕ. The Wald's test in (32) for our methods is used for testing the significance with the coefficient β_1 estimated from the MLE $\hat{\beta}_1 = \frac{\bar{y}_1}{w\bar{y}_0}$ and the variance estimated from (34) under the two different approaches. The fold change between the two groups is considered significantly different when the 2-sided p-value ≤ 0.05. We repeated the procedure 5000 times, and the power was calculated as the percentage of time that the H_0 was rejected in the 5000 simulated data. The results are listed in Table 7 and 8.

Table 7: Sample size (n_0) and simulated power in parentheses from three methods given a balanced design k = 1, an equal read depth w = 1 and an 80% nominal power.

ρ	ϕ	u_0	n_0 (Simulated power) & $\alpha = 0.05$			n_0 (Simulated power) & $\alpha^* = 0.000425$			Simulated power & $\alpha^* = 0.000425$	
			M1	M2	M3	M1	M2	M1	M2: $n_0 = n_{M1}$	M3: $n_0 = n_{M1}$
0.5	0.1	1	52(.78)	63(.77)	54(.81)	127(.79)	159(.79)	124(.80)	.62	.83
		5	13(.79)	15(.79)	14(.82)	32(.79)	38(.79)	32(.82)	.66	.82
		10	8(.78)	9(.77)	9(.83)	20(.78)	23(.78)	20(.80)	.67	.80
		50	4(.77)	4(.75)	5(.77)	10(.77)	11(.79)	11(.83)	.72	.77
	0.5	1	65(.78)	77(.78)	68(.80)	159(.78)	190(.78)	157(.80)	.66	.81
		5	26(.79)	28(.78)	27(.81)	63(.78)	70(.78)	64(.80)	.72	.79
		10	21(.78)	22(.78)	22(.80)	52(.80)	55(.80)	53(.81)	.75	.80
		50	17(.79)	18(.80)	18(.81)	42(.80)	43(.81)	43(.81)	.79	.80
	1	1	82(.80)	93(.79)	84(.81)	198(.79)	230(.79)	198(.81)	.68	.81
		5	42(.78)	45(.77)	44(.80)	103(.80)	110(.81)	105(.81)	.75	.81
		10	38(.81)	39(.80)	39(.81)	91(.79)	94(.79)	93(.80)	.77	.79
		50	34(.80)	34(.80)	35(.81)	82(.80)	82(.79)	84(.81)	.79	.79
0.8	0.1	1	386(.80)	414(.79)	396(.81)	938(.79)	1015(.79)	946(.80)	.74	.80
		5	102(.80)	108(.79)	105(.81)	249(.79)	264(.79)	251(.80)	.75	.80
		10	67(.80)	70(.80)	69(.80)	163(.79)	170(.79)	164(.80)	.76	.80
		50	39(.80)	39(.80)	40(.81)	94(.80)	95(.80)	95(.81)	.79	.80
	0.5	1	512(.79)	540(.78)	532(.80)	1244(.79)	1321(.80)	1254(.80)	.75	.79
		5	229(.80)	234(.79)	232(.80)	555(.79)	570(.80)	559(.80)	.78	.79
		10	193(.80)	196(.79)	196(.80)	469(.79)	477(.79)	472(.80)	.78	.79
		50	165(.80)	165(.80)	166(.81)	400(.80)	402(.80)	402(.80)	.79	.80
	1	1	670(.81)	687(.80)	682(.81)	1627(.79)	1704(.79)	1639(.80)	.76	.79
		5	386(.80)	392(.80)	391(.81)	938(.80)	953(.80)	944(.81)	.79	.80
		10	351(.79)	353(.79)	354(.80)	852(.79)	859(.79)	857(.79)	.78	.79
		50	322(.80)	323(.80)	325(.81)	783(.80)	784(.80)	787(.81)	.80	.80
1.2	0.1	1	480(/81)	452(.81)	491(.80)	1166(.81)	1088(.81)	1175(.80)	.85	.79
		5	134(.80)	128(.80)	137(.80)	325(.81)	309(.81)	328(.81)	.84	.80
		10	91(.80)	88(.80)	92(.79)	220(.80)	212(.80)	222(.80)	.82	.80
		50	56(.79)	55(.79)	57(.80)	136(.81)	134(.80)	137(.81)	.82	.81
	0.5	1	669(.81)	641(.81)	681(.81)	1624(.80)	1547(.80)	1636(.80)	.83	.79
		5	323(.81)	317(.81)	327(.81)	784(.80)	768(.81)	788(.81)	.82	.80
		10	279(.80)	277(.81)	282(.80)	678(.81)	671(.81)	682(.81)	.81	.81
		50	245(.81)	244(.81)	247(.81)	594(.81)	593(.81)	598(.82)	.81	.81
	1	1	905(.81)	878(.81)	919(.80)	2198(.80)	2120(.80)	2213(.80)	.82	.79
		5	559(.80)	553(.80)	564(.81)	1357(.80)	1341(.80)	1364(.80)	.81	.79
		10	516(.81)	513(.81)	520(.81)	1252(.80)	1244(.80)	1258(.80)	.80	.79
		50	481(.81)	480(.81)	485(.81)	1168(.80)	1166(.80)	1173(.80)	.80	.80
1.5	0.1	1	89(.81)	78(.81)	92(.80)	216(.82)	185(.83)	218(.80)	.90	.80
		5	25(.82)	23(.82)	27(.82)	62(.82)	56(.82)	63(.81)	.88	.80
		10	18(.82)	16(.80)	18(.81)	43(.82)	39(.80)	43(.80)	.86	.80
		50	11(.79)	11(.79)	12(.82)	27(.79)	26(.78)	28(.81)	.81	.79
	0.5	1	127(.81)	116(.81)	131(.80)	309(.81)	278(.82)	311(.80)	.87	.80
		5	64(.81)	61(.80)	65(.80)	155(.81)	148(.81)	156(.81)	.83	.80
		10	56(.81)	55(.81)	57(.81)	135(.80)	132(.80)	137(.81)	.82	.80
		50	49(.80)	49(.80)	50(.80)	120(.80)	119(.80)	121(.81)	.81	.80
	1	1	175(.81)	164(.81)	180(.80)	425(.80)	394(.81)	429(.80)	.85	.79
		5	111(.79)	109(.80)	114(.80)	270(.80)	264(.80)	273(.80)	.82	.80
		10	103(.79)	102(.80)	105(.80)	251(.81)	248(.81)	254(.81)	.82	.81
		50	97(.80)	97(.81)	99(.81)	236(.81)	235(.81)	238(.81)	.81	.80
2.0	0.1	1	28(.84)	22(.84)	29(.82)	67(.83)	51(.84)	67(.81)	.94	.81
		5	8(.82)	7(.82)	9(.82)	20(.84)	17(.84)	20(.82)	.92	.82
		10	6(.84)	5(.79)	6(.81)	14(.83)	12(.81)	14(.81)	.89	.81
		50	4(.83)	4(.84)	4(.82)	9(.80)	9(.82)	10(.84)	.82	.79
	0.5	1	41(.82)	35(.82)	42(.80)	99(.82)	83(.83)	99(.80)	.90	.80
		5	21(.81)	20(.81)	22(.81)	52(.82)	48(.81)	53(.82)	.85	.81
		10	19(.81)	18(.80)	20(.82)	46(.81)	44(.80)	47(.82)	.83	.80

		50	17(.82)	17(.82)	18(.83)	41(.80)	41(.81)	42(.81)	.81	.80
	1	1	57(.81)	51(.81)	59(.80)	139(.81)	123(.81)	140(.80)	.87	.79
		5	38(.81)	36(.80)	39(.81)	91(.80)	88(.80)	93(.80)	.82	.79
		10	35(.81)	35(.82)	36(.81)	85(.81)	84(.81)	87(.81)	.82	.80
		50	33(.80)	33(.80)	34(.80)	81(.81)	80(.80)	82(.81)	.81	.80
3.0	0.1	1	10(.87)	7(.86)	10(.81)	24(.87)	15(.85)	23(.83)	.98	.85
		5	3(.84)	2(.77)	4(.89)	7(.80)	6(.86)	8(.87)	.93	.78
		10	2(.79)	2(.84)	3(.91)	5(.78)	4(.76)	6(.88)	.88	.75
		50	1(.65)	1(.67)	2(.90)	4(.87)	3(.72)	4(.85)	.89	.85
	0.5	1	15(.64)	12(.82)	16(.83)	37(.84)	28(.82)	37(.82)	.94	.82
		5	8(.78)	8(.82)	9(.82)	20(.82)	18(.80)	21(.83)	.86	.79
		10	7(.77)	7(.79)	8(.81)	18(.81)	17(.80)	19(.82)	.83	.78
		50	7(.83)	7(.83)	7(.81)	16(.80)	16(.80)	17(.82)	.80	.78
	1	1	22(.83)	19(.82)	23(.82)	53(.83)	44(.82)	53(.81)	.91	.81
		5	15(.81)	14(.79)	16(.82)	36(.81)	34(.81)	37(.80)	.84	.79
		10	14(.81)	14(.82)	15(.81)	34(.81)	33(.81)	35(.81)	.82	.79
		50	13(.80)	13(.80)	14(.81)	127(.79)	159(.79)	124(.80)	.81	.78

Table 8: Sample size (n_0) and simulated power in parentheses from our two methods given an imbalanced design k=3/2, an equal read depth w=1, an 80% nominal power and significance level of $\alpha = 0.05$ for testing single gene.

ρ	ϕ	u_0	n_0 (Simulated power %)	
			M1	**M2**
0.5	0.1	1	41(0.83)	52(.78)
		5	10(.83)	13(.79)
		10	7(.84)	8(.80)
		50	3(.74)	4(.81)
	0.5	1	52(.83)	63(.77)
		5	21(.80)	23(.78)
		10	17(.79)	19(.79)
		50	14(.78)	15(.80)
	1	1	65(.81)	77(.79)
		5	35(.79)	37(.78)
		10	31(.80)	32(.79)
		50	28(.80)	28(.79)
0.8	0.1	1	315(.84)	343(.78)
		5	84(.83)	90(.79)
		10	55(.82)	58(.79)
		50	32(.80)	33(.80)
	0.5	1	420(.83)	448(.80)
		5	189(.81)	195(.80)
		10	160(.80)	163(.79)
		50	137(.80)	138(.80)
	1	1	552(.83)	579(.79)
		5	321(.81)	326(.81)
		10	292(.79)	294(.79)
		50	268(.79)	269(.79)
1.2	0.1	1	407(.86)	379(.81)
		5	113(.84)	107(.81)
		10	76(.82)	73(.80)
		50	47(.81)	46(.79)
	0.5	1	564(.84)	536(.81)
		5	270(.82)	265(.81)
		10	233(.81)	231(.81)

		50	204(.81)	204(.81)
	1	1	761(.84)	733(.80)
		5	467(.82)	461(.80)
		10	430(.81)	428(.81)
		50	401(.81)	400(.81)
1.5	0.1	1	77(.86)	66(.83)
		5	22(.85)	19(.82)
		10	15(.85)	14(.83)
		50	9(.79)	9(.79)
	0.5	1	109(.85)	97(.81)
		5	54(.83)	51(.81)
		10	47(.81)	46(.81)
		50	41(.80)	41(.80)
	1	1	149(.84)	137(.81)
		5	93(.80)	91(.80)
		10	86(.80)	85(.81)
		50	81(.81)	81(.81)
2.0	0.1	1	25(.89)	19(.85)
		5	7(.85)	6(.82)
		10	5(.83)	4(.79)
		50	3(.79)	3(.80)
	0.5	1	35(.85)	30(.83)
		5	18(.83)	17(.83)
		10	16(.83)	15(.81)
		50	14(.83)	14(.81)
	1	1	49(.83)	43(.82)
		5	32(.82)	30(.81)
		10	29(.81)	29(.82)
		50	28(.81)	28(.81)
3.0	0.1	1	9(.90)	6(.87)
		5	3(.90)	2(.84)
		10	2(.87)	2(.90)
		50	1(.71)	1(.72)
	0.5	1	13(.86)	10(.83)
		5	7(.81)	6(.79)
		10	6(.80)	6(.82)
		50	6(.84)	6(.84)
	1	1	19(.84)	16(.83)
		5	12(.79)	12(.81)
		10	12(.83)	11(.79)
		50	11(.81)	11(.81)

Results: Table 7 illustrates the sample size and simulated power in parentheses for the above settings with an equal read depth $w = 1$ under a balanced design $k = 1$ in the case of testing a single gene. We first examined the changing pattern of the sample size with different parameter values of u_0, ϕ and ρ for all three methods. Table 7 shows that the sample size n in all the methods decreases as read counts u_0 increases from 1 to 50 given a fixed ρ and ϕ or vice versa, which is expected. This pattern suggests that a larger sample size for a lowly expressed gene is required in order to achieve an empirical power close to 80% in a DEG analysis. Given fixed u_0 and ρ, the sample size increases as ϕ increases. This is also expected due to higher variation of the genes across samples. Moreover, given fixed ϕ and u_0, the sample size decreases as ρ increases. The results suggest that a smaller n is required for a larger fold change of a gene between two conditions. Similarly, Table 8 for testing a single gene illustrates the sample size and simulated power (in parentheses) for the above settings in case of an imbalanced design with $k = 3/2$ and equal read depth $w = 1$. In this situation, a smaller sample size for the control group is required comparing with a balanced design in Table 7. But a larger sample size for the treatment group is required in an imbalanced design.

The performance of these methods is further evaluated by its accuracy, which is the absolute difference between the empirical power and nominal power (80%). The better sample size method should have the empirical power close to the nominal power suggested in the previous study [39]. Overall, samples sizes estimated from our two methods and the existing method vary in different parameter settings. The corresponding power calculated from the simulation studies is close to the nominal power 80% in most cases for all the methods. However, we found that the empirical power estimated by method one (M1) and method three (M3, the existing method) appear to be closer to an 80% nominal power than that achieved by method two (M2). Although M2 trails to the other methods in some settings, it appears overestimating the sample size when a fold change is less than one, and it underestimates the sample size when a fold change is greater than one. A further comparison among these methods will be discussed when we test multiple genes.

3.2. Sample size and power estimation in the case of testing multiple genes with FDR-controlling

In this simulation, the total number of genes W in RNA-seq data is set to be 10,000; the true DEG W_1 is set to be 100, and $W_0 = W - W_1$ is the number of genes not differentially expressed under H_0. The expected number of true DEGs and the empirical power corresponding to the nominal power 80% will be estimated. The rest parameter settings including u_0, ρ, w, ρ, ϕ and k are similar to those for testing a single gene. In addition, we set $= 1$ or 1.2 and $k = \frac{1}{3}, 1$ or $\frac{3}{2}$. With these settings, a significance level α^* in (20) is calculated as 0.000425 given a nominal FDR $(f = 0.05)$.

For each combination of these parameter settings, sample sizes were calculated using the formulas in (43-44), and the existing method given an 80% nominal power and a two-sided test at $\alpha^* = 0.000425$, respectively. The corresponding power from the calculated sample size was obtained via a simulation using a two-sided Wald's test statistic in (32).

Given the estimated sample size and each designed parameter setting, an empirical power was calculated from three methods via simulation, using similar procedure described for testing a single gene. A fold change between two treatment groups for the multiple corrections is considered to be significantly different when a two-sided p-value is $\leq \alpha^* = 0.000425$, which is much smaller than the nominal significance level $\boldsymbol{\alpha} = 0.05$ for testing a single gene. The empirical power was obtained as the percentage of the number of times that the null hypothesis is rejected at the significance level α^* in the 5000 simulated data.

Results for each combination of the desired parameters with different read depth w and ratio of sample size k in the case of balanced or imbalanced design will be discussed in details.

Results: The sample size and simulated power in parentheses in the case of a balanced design $k = 1$ with equal read depth $w = 1$ when testing multiple genes are also listed in Table 7. We observed that the changing pattern of the sample size varying for different parameter values of u_0, ϕ and ρ in the three methods (Table 7) is similar to that in the case of testing a single gene. However, given a similar setting, a much larger sample size is required for testing multiple genes compared to a single gene. Since the empirical power varies for the corresponding sample sizes from different methods, a simulated power for each method is further calculated given a fixed sample size n_{M1} from M1 (Table 7). We observed that M1 and M3 can achieve a higher empirical power close to 80% in most settings when the fold change is 0.5 and 0.8. However, M2 can achieve a higher power in most settings when the fold change is greater than 1. A further comparison of these methods is evaluated by the difference between the simulated power and the 80% nominal power. A non-parametric Wilcoxon test and a parametric T-test statistic were used for testing the significance, respectively. The results from both test statistics (Table 9) indicate that the estimated power between M1 and M3 is not significantly different with a p-value greater than 0.05. However, both of these methods are significantly different from M2 with p-values less than 0.01.

Table 10 represents the sample size and simulated power in parentheses from our two methods in the case of an imbalanced design $k = 3/2$ with read depth $w = 1$ and $w = 1.2$, respectively. Given the same settings and $k = 2/3$, the sample size and simulated power in parentheses are listed in Table 11. We observed that the required sample size with unequal size factors given $w = 1.2$ is slightly smaller than that with an equal size factors (Tables 10 and 11).

Table 9: p-values obtained from a paired Wilcox test and t-test given the absolute value of difference between simulated power and an 80% nominal power for each method.

Wilcox test		M2	M3.exact
	M1	<0.0001	0.280
	M2		<0.0001
T test			
	M1	<0.0001	0.267
	M2		<0.0001

Table 10: Sample sizes and simulated power in parentheses from two methods for testing multiple genes given an unequal sample size k = 3/2, an 80% nominal power, a 0.05 FDR and the significant level α^*= 0.000425.

ρ	ϕ	u_0	n_0 (Simulated power) & $w = 1$		n_0 (Simulated power) & $w = 1.2$	
			M1	M2	M1	M2
0.5	0.1	1	99(.77)	131(.78)	90(.77)	120(.77)
		5	25(.77)	31(.76)	23(.76)	29(.76)
		10	16(.77)	19(.76)	15(.76)	18(.76)
		50	8(.75)	9(.76)	8(.77)	9(.78)
	0.5	1	126(.77)	157(.77)	117(.78)	147(.78)
		5	52(.78)	58(.78)	50(.79)	56(.78)
		10	42(.78)	45(.77)	41(.78)	44(.78)
		50	35(.79)	36(.80)	35(.80)	35(.79)
	1	1	159(.79)	190(.77)	150(.79)	180(.78)
		5	85(.80)	91(.79)	83(.80)	89(.80)
		10	75(.79)	79(.79)	74(.79)	77(.79)
		50	68(.79)	69(.80)	68(.80)	68(.79)
0.8	0.1	1	765(.79)	842(.80)	712(.79)	778(.79)
		5	204(.79)	220(.79)	193(.79)	207(.80)
		10	134(.79)	142(.79)	129(.79)	135(.79)
		50	78(.80)	79(.80)	77(.80)	78(.80)
	0.5	1	1021(.79)	1098(.79)	967(.79)	1034(.79)
		5	459(.80)	475(.79)	449(.79)	462(.79)
		10	389(.79)	397(.80)	384(.80)	390(.79)
		50	333(.79)	335(.80)	332(.79)	333(.79)
	1	1	1340(.79)	1417(.79)	1286(.79)	1353(.79)
		5	778(.80)	794(.80)	768(.80)	781(.80)
		10	708(.79)	716(.79)	703(.79)	709(.79)
		50	652(.79)	653(.79)	651(.79)	652(.79)

1.2	0.1	1	987(.81)	910(.81)	934(.82)	846(.81)
		5	274(.82)	258(.82)	263(.81)	246(.81)
		10	185(.81)	177(.81)	179(.80)	171(.80)
		50	113(.80)	112(.80)	112(.80)	111(.80)
	0.5	1	1370(.80)	1292(.80)	1316(.80)	1228(.80)
		5	656(.81)	641(.81)	646(.82)	628(.81)
		10	567(.81)	559(.81)	562(.81)	553(.81)
		50	496(.81)	494(.81)	495(.82)	493(.81)
	1	1	1847(.80)	1770(.80)	1794(.80)	1706(.80)
		5	1134(.81)	1118(.80)	1123(.80)	1106(.81)
		10	1045(.81)	1037(.81)	1039(.80)	1031(.81)
		50	973(.80)	972(.80)	972(.80)	971(.80)
1.5	0.1	1	187(.83)	155(.83)	178(.83)	145(.83)
		5	53(.82)	47(.83)	51(.82)	44(.82)
		10	36(.81)	33(.81)	35(.82)	32(.81)
		50	23(.81)	22(.80)	22(.79)	22(.81)
	0.5	1	264(.82)	233(.82)	255(.82)	222(.82)
		5	130(.81)	124(.81)	128(.81)	122(.81)
		10	113(.81)	110(.81)	112(.81)	109(.81)
		50	100(.80)	99(.80)	100(.80)	99(.80)
	1	1	361(.81)	329(.81)	352(.82)	319(.81)
		5	227(.81)	220(.80)	225(.81)	218(.80)
		10	210(.81)	207(.81)	209(.81)	206(.81)
		50	197(.81)	196(.81)	196(.81)	196(.81)
2.0	0.1	1	59(.84)	43(.84)	57(.85)	40(.84)
		5	17(.83)	14(.84)	17(.85)	13(.82)
		10	12(.83)	10(.82)	12(.84)	10(.83)
		50	10(.93)	8(.85)	8(.84)	7(.79)
	0.5	1	86(.83)	70(.83)	84(.84)	67(.83)
		5	44(.82)	40(.80)	43(.81)	40(.81)
		10	38(.80)	37(.81)	38(.81)	36(.79)
		50	34(.81)	34(.81)	34(.80)	34(.81)
	1	1	119(.82)	103(.82)	117(.82)	100(.81)
		5	77(.81)	73(.80)	76(.81)	73(.81)
		10	71(.80)	70(.81)	71(.80)	70(.81)
		50	67(.81)	67(.81)	67(.81)	67(.81)
3.0	0.1	1	22(.87)	13(.87)	21(.89)	12(.85)
		5	6(.80)	5(.86)	6(.81)	5 (.87)
		10	5(.88)	4(.87)	5(.88)	4(.88)
		50	4(.94)	3(.83)	3(.80)	3(.83)
	0.5	1	32(.85)	24(.84)	32(.86)	23(.83)
		5	17(.82)	15(.80)	17(.83)	15(.81)
		10	15(.82)	14(.80)	15(.82)	14(.81)
		50	14(.83)	13(.79)	14(.83)	13(.79)
	1	1	46(.84)	37(.83)	45(.84)	36(.81)
		5	30(.81)	28(.80)	30(.82)	28(.81)
		10	28(.82)	27(.81)	28(.81)	27(.81)
		50	27(.81)	27(.82)	27(.81)	27(.82)

Table 11: Sample size (n_0) and simulated power in parentheses from our two methods given k = 2/3, an 80% nominal power, the significant level α^* of 0.000425 and equal read depth or unequal read depth of w.

ρ	ϕ	u_0	n_0 (Simulated power) & $w = 1$		n_0 (Simulated power %) & $w = 1.2$	
			M1	M2	M1	M2
0.5	0.1	1	169(.80)	200(.79)	149(.79)	183(.79)
		5	42(.80)	48(.79)	38(.80)	45(.80)
		10	26(.79)	29(.78)	24(.79)	27(.79)
		50	13(.79)	14(.81)	13(.81)	13(.78)
	0.5	1	208(.79)	240(.79)	188(.78)	223(.79)
		5	81(.79)	88(.80)	77(.80)	84(.79)
		10	65(.78)	69(.79)	63(.79)	69(.81)
		50	53(.82)	53(.81)	52(.81)	53(.80)
	1	1	258(.79)	290(.79)	238(.79)	272(.79)
		5	131(.80)	137(.80)	127(.80)	134(.80)
		10	115(.80)	118(.80)	113(.80)	117(.80)
		50	102(.79)	103(.81)	102(.81)	103(.80)
0.8	0.1	1	1196(.80)	1273(.80)	1076(.80)	1169(.79)
		5	316(.79)	331(.80)	292(.80)	310(.79)
		10	206(.78)	213(.79)	194(.80)	203(.79)
		50	118(.80)	119(.80)	115(.81)	117(.80)
	0.5	1	1579(.81)	1656(.80)	1459(.80)	1552(.80)
		5	698(.79)	714(.79)	675(.80)	693(.80)
		10	588(.79)	596(.80)	576(.79)	596(.80)
		50	500(.80)	502(.80)	498(.79)	500(.79)
	1	1	2057(.80)	2134(.80)	1938(.80)	2031(.81)
		5	1177(.80)	1192(.79)	1153(.80)	1172(.80)
		10	1067(.79)	1075(.80)	1055(.80)	1064(.81)
		50	979(.79)	980(.79)	976(.79)	978(.79)
1.2	0.1	1	1433(.81)	1356(.81)	1314(.81)	1253(.80)
		5	401(.81)	386(.81)	377(.80)	365(.81)
		10	272(.81)	265(.81)	260(.80)	254(.80)
		50	169(.80)	168(.81)	167(.80)	166(.80)
	0.5	1	2007(.80)	1929(.81)	1887(.80)	1826(.80)
		5	975(.79)	959(.80)	951(.80)	939(.80)
		10	846(.81)	838(.81)	834(.80)	838(.81)
		50	742(.80)	741(.80)	740(.80)	739(.80)
	1	1	2723(.80)	2646(.80)	2604(.80)	2543(.80)
		5	1691(.80)	1676(.80)	1667(.79)	1655(.79)
		10	1562(.80)	1555(.80)	1550(.80)	1544(.80)
		50	1459(.79)	1458(.79)	1457(.80)	1455(.80)
1.5	0.1	1	261(.82)	229(.81)	241(.82)	213(.81)
		5	75(.81)	69(.81)	71(.81)	66(.81)

		10	52(.80)	49(.81)	50(.82)	47(.80)
		50	34(.80)	33(.79)	33(.79)	33(.80)
	0.5	1	377(.81)	345(.82)	357(.81)	329(.82)
		5	191(.81)	185(.81)	187(.80)	182(.81)
		10	168(.80)	165(.80)	166(.80)	165(.82)
		50	150(.80)	149(.80)	149(.80)	149(.80)
	1	1	522(.80)	490(.80)	502(.80)	474(.80)
		5	336(.79)	330(.79)	332(.80)	327(.80)
		10	313(.80)	310(.80)	311(.80)	308(.80)
		50	294(.80)	294(.80)	294(.80)	293(.80)
2.0	0.1	1	79(.84)	63(.84)	74(.83)	59(.83)
		5	24(.82)	21(.83)	23(.83)	20(.82)
		10	17(.81)	15(.81)	16(.80)	15(.83)
		50	11(.77)	11(.79)	11(.78)	11(.80)
	0.5	1	119(.83)	103(.82)	114(.82)	99(.82)
		5	63(.80)	60(.81)	62(.80)	59(.80)
		10	57(.81)	55(.80)	56(.80)	55(.81)
		50	51(.79)	51(.80)	51(.79)	51(.80)
	1	1	169(.82)	152(.81)	164(.82)	148(.81)
		5	113(.80)	110(.80)	112(.80)	109(.80)
		10	106(.80)	105(.80)	106(.80)	104(.80)
		50	101(.80)	100(.79)	100(.80)	100(.80)
3.0	0.1	1	28(.87)	19(.85)	26(.85)	18(.85)
		5	9(.84)	7(.81)	8(.79)	7(.85)
		10	6(.78)	5(.75)	6(.80)	5(.76)
		50	4(.71)	4(.74)	4(.72)	4(.74)
	0.5	1	43(.81)	35(.81)	42(.82)	34(.81)
		5	24(.79)	23(.81)	24(.81)	23(.82)
		10	22(.81)	21(.80)	22(.81)	21(.80)
		50	20(.79)	20(.80)	20(.80)	20(.80)
	1	1	63(.81)	55(.82)	62(.82)	53(.81)
		5	44(.80)	43(.81)	44(.80)	42(.80)
		10	42(.80)	41(.80)	42(.80)	41(.80)
		50	40(.79)	40(.79)	40(.79)	40(.79)

4. An example of real RNA-seq dataset

In this example, we illustrate how to estimate the sample size using formulas (21-22) for testing multiple genes in a real study. The RNA-seq dataset containing forty-one Triple Negative Breast Cancer (TNBC) primary tumors and twenty-one uninvolved breast tissue samples that were adjacent to TNBC primary tumors in .fastq format were downloaded from NCBI GEO (series ID GSE58135) [28]. The raw sequencing files were mapped to the human hg19 reference genome using tophat2 (v2.0.13) with bowtie version (2.2.3.0). The mapped counts for 57,773 genes per sample were then extracted using HTSeq-scripts-count (version 2.7). Three of the forty-two TNBC samples failed during the process of mapping or extraction of the read counts. Therefore, a total of sixty samples (21 normal and 39 TNBC samples) are used in this application. After filtering the genes with the mean read counts less than one in two groups, the sixty samples containing 36,762 genes were loaded into *edgeR* for estimating the parameters of common dispersion and size factors. We used the "TMM" normalization method to estimate the normalization factors called size factors [17]. The mean size factors of $\bar{s}_0$ in the normal group and $\bar{s}_1$ in the cancer group are 0.936 and 1.065, respectively. Their ratio (w) is 1.14. The estimated common dispersion ϕ is 0.49 [22].

We assumed the top 500 of 36,762 genes (1.4%) are prognostic and have the largest fold changes for up-regulated or down-regulated genes. The minimum of average read counts among these genes in the control group served as the pilot data and was estimated as $u_0 \approx 3$ [16]. In addition, the sample sizes were estimated by setting $u_0 = 10,\ 20$ and 50. Suppose we want to set the nominal power to be 80%, which indicates we want to identify 400 or more of the prognostic genes. Under the control FDR at $f = 0.05$ or 0.01, and 80% power, we can set $\mathrm{t} = 36{,}762$, $t_1 = 500$, $t_0 = \mathrm{t} - t_1$ and $r_1 = 400$. The parameters ρ_g $(g = 1, \dots, 36{,}762)$ were assumed to be unknown. We set the mean counts in the control group to be $u_{0g} = 3, 10$, 20 or 50 and FC to be $\rho_g = 0.5, 1.5, 2$ or 3 with common dispersion $\phi_g = 0.49$. With these settings, the new $\alpha^* = 5.81 \times 10^{-4}$ or 1.23×10^{-3} was obtained from equation (14) at a desired

FDR ($f = 0.10$ or 0.05). Finally, the sample sizes were estimated by substituting α^* and the power into equations (43-44) for each method (Table 12).

Table 12: Sample sizes estimated from two methods for a breast cancer RNA-seq data given *k*=1.9, *w* = 1.14, $\phi = 0.49$, an 80% nominal power and the significant level α^*.

ρ	$\boldsymbol{u_0}$	**Sample size($\boldsymbol{n_0}$) FDR = 0.05 & $\boldsymbol{\alpha^* = 0.000581}$**		**Sample size($\boldsymbol{n_0}$) FDR = 0.10 & $\boldsymbol{\alpha^* = 0.001226}$**	
		$\boldsymbol{n_{M1}}$	$\boldsymbol{n_{M2}}$	$\boldsymbol{n_{M1}}$	$\boldsymbol{n_{M2}}$
0.5	3	53	63	48	57
	10	36	39	32	35
	20	32	34	29	30
	50	30	31	27	28
0.8	3	469	490	425	443
	10	333	340	302	308
	20	304	308	276	278
	50	287	288	260	261
1.2	3	667	639	604	579
	10	489	480	443	435
	20	451	447	408	404
	50	428	426	387	386
1.5	3	132	121	120	110
	10	98	95	89	86
	20	91	89	82	81
	50	86	86	78	78
2.0	3	44	39	40	35
	10	33	32	30	29
	20	31	30	28	27
	50	29	29	27	28
3.0	3	17	14	16	13
	10	13	12	12	11
	20	12	12	11	11
	50	12	12	11	10

Table 12 reports the sample sizes n_0 in the control group and $n_1 = kn_0$ in the cancer group under different settings for an imbalanced design with $k = 1.9$. We found that the original RNA-seq experiment [28] with a minimum $n_0 = 17$ and $n_1 = 33$ from method one (M1) can detect more than 80% of the prognostic genes at a two-sided FDR (f = 0.05) and the minimum $u_0 = 3$ if the desired fold change for up-regulated genes is 3 or more. For an imbalanced experimental design, a new study needs 53 normal control and 101 cancer patient samples from M1 to achieve the detection power at 80% if the desired fold change for the down-regulated gene is 2 or more. With the same settings, a new study needs 44 normal control and 84 cancer patient samples for the up-regulated genes with a desired FC of 2 or more, which is fewer than the down-regulated genes. As the desired FC decreases to 1.5 or 1.2, a new study in each group requires a larger number of patients in order to achieve an 80% nominal power at a two-sided test with a 0.05 FDR. Moreover, Table 12 also shows that a smaller sample size for all settings is required if the FDR increases from 0.05 to 0.10.

5. Discussion and conclusion

In this study, we used a GLM and a two-sided Wald's test statistic of the model parameter β_1 to derive two novel sample size calculation methods (n_{M1} and n_{M2}) for RNA-seq data. We incorporated a common dispersion parameter and the size factors via a log link function in the GLM with a NB distribution. The variance of the Wald's test was estimated from the variance-covariance matrix with the parameters that are estimated from the MLE and CMLE. Since the dispersion parameter has no closed form for the MLE, it is difficult to explicitly derive a sample size formula using a Wald's test statistic. Therefore, in this study, we assume that the dispersion parameter is a known constant value while deriving the variance from the model parameters. This alternative choice has been commonly used by previous studies when deriving the sample size from a NB distribution [37, 39].

Previous sample size calculations for RNA-seq data focus on a balanced design only. In this paper, sample size calculation methods are presented under different settings with either a balanced or an imbalanced design at a two-sided test, which is applicable in many situations. Comparing with previous studies in identifying DEG, the cutoff value for the minimum average read counts per gene is commonly chosen to be one or five, while we used a wide range of gene read counts in the control group (1, 5, 10, 20 and 50) with a minimum of one read in the case of lowly expressed genes in the simulation. For a lowly expressed gene, a larger sample size is needed according to our simulation. However, a study requiring a large sample size to achieve a nominal power at 80% or higher is not feasible in practice due to the cost. As an alternative, the higher read depth sequencing may be chosen to increase the mean read counts in the sample, instead of directly increasing the sample size. Moreover, when testing multiple genes in the simulation, we arbitrarily chose 10, 000 genes and 10% true DEGs. In reality, the total number of detected genes, at least in RNA-seq data, could vary depending upon the read depth in each sequencing sample. One of the limitations in this study is that we derived sample size formulas for two groups. In future studies, we will consider the case of multiple treatment groups and paired samples.

In summary, sample sizes estimated from our two methods and the existing method vary in different parameter settings. The corresponding power calculated from the simulation studies is close to the 80% nominal power in most cases. We found that our method M1 and the existing method M3 appear to be similar in most settings. Since our sample size formulas are based on the Wald's test statistic that utilizes a large sample property of the MLE, it may not be an appropriate test for a small sample size. In contrast, the M3 from the exact test may be a better choice for an experimental design with a small sample size. However, our method can provide an alternative choice for researchers to quickly and roughly estimate sample sizes when designing a RNA-seq experiment for clinical research studies.

Finally, this work with all the contents and tables was submitted into Biostatistics in Medicine.

NORMALIZATION METHODS FOR RNA-SEQ DATA

1. Introduction

High-throughput RNA sequencing (RNA-seq) has become the preferred choice for gene expression studies, especially as the cost of high-quality sequencing has decreased due to technological improvements. These improvements have enabled transcriptome studies with a large range of applications including the identification of alternatively spliced isoforms [6, 15, 41], *de novo* assembly of transcripts to identify novel genes and isoforms [4, 42, 43], the detection of single-nucleotide polymorphisms (SNPs) [44, 45] and new single nucleotide variants (SNVs) [46], and a characterization of mRNA editing [47]. In addition, RNA-seq enables the detection of rare transcripts while allowing for high coverage of the genome, which cannot be identified as well by microarray technology [48]. However, the most common and popular application for RNA-seq experiments is the identification of differentially expressed genes (DEGs) between conditions or tissues. DEGs may serve as disease biomarkers that are helpful in clinical diagnosis, with possible implications for prevention, prognosis and treatment [49, 50]. For example, data derived from sequencing can be helpful in predicting clinical outcomes and for discovering subtypes of breast cancers [9, 51].

Currently, there are a plethora of sequencing platforms with new generations being developed on a regular basis. All platforms require similar sample pre-processing and subsequent analytical steps, which are summarized in RNA-seq workflows described by Zeng *et al*. [52]. Briefly, the mRNA-seq workflow consists of four major steps including: 1) RNA-seq library construction; 2) sequencing and mapping; and 3) normalization and statistical modeling to identify the DEGs or transcript isoforms; and 4) Exploration of biological insights including biological functions and pathways of the DEGs or their isoforms. Figure 3 adapted and modified from the reference figure[52] illustrates the procedure of step 1 and step 2; and Figure 4 illustrates the procedure of the analysis of RNA-seq in step 3 and step 4.

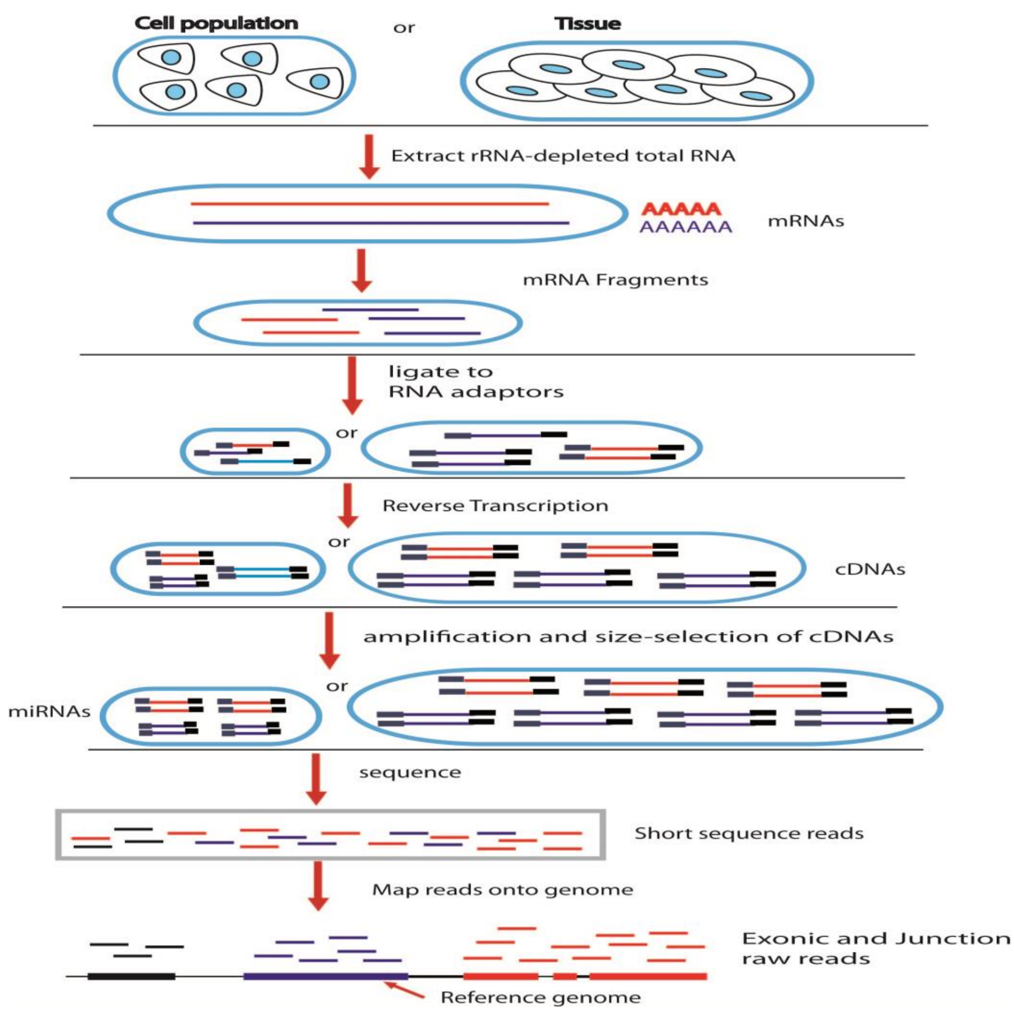

Figure 3: mRNA-seq workflow illustrates the procedures from the library preparation to the sequencing mapping.

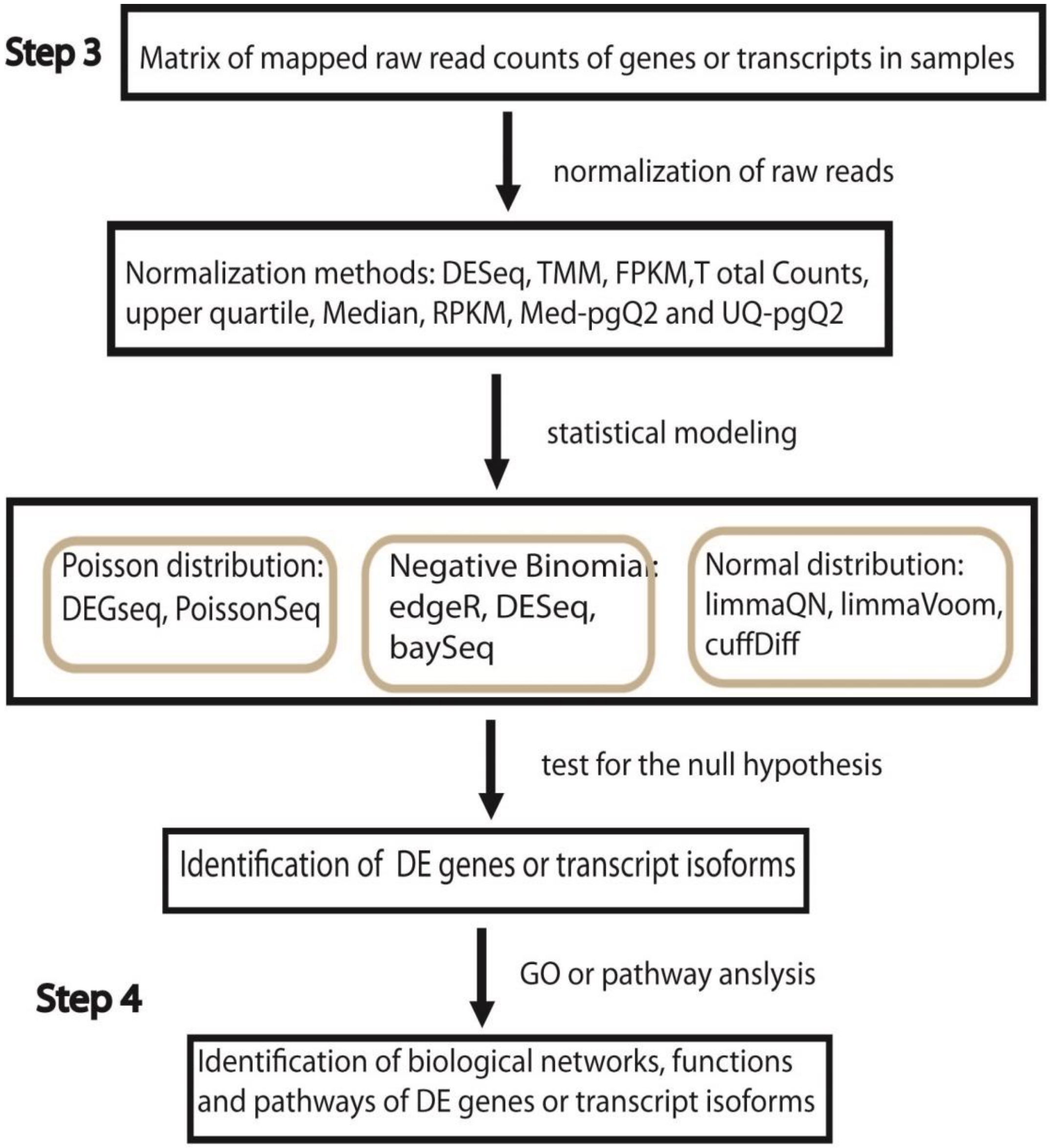

Figure 4: RNA-seq workflow illustrates the procedure of normalization, statistical test for identify DEGs and biological functions.

Following the second step of the RNA-seq workflow in Figure 3, raw mapped reads generated by an aligner such as TopHat2 [53] or STAR [54] for splicing alignment or *SOAP2* for short read alignment [55] are further normalized by a variety of methods, which generally include within and between library normalization. Normalization is a key step for the identification of DEGs as shown by many gene expression studies for both microarray and RNA-seq data [36, 56-58].

Prior to the advent of next generation sequencing (NGS), high-throughput microarray data had been widely used for gene expression studies. Use of this technique induced several systematic non-biological variables which obfuscated data analysis. These variables included unequal quantities of starting RNA, differences in hybridization between chips and differences between manufactured chips. In order to minimize these variables to reveal true biological differences, the normalization of one-color array and two-color microarrays became essential. This allowed experiments to be standardized so that relative gene expression levels could be compared between chips [59]. The basic normalization steps in a one-color microarray included data transformation, per chip normalization and per gene/spot normalization. Per chip normalization (global normalization) is used to correct systematic variation in which the intensities of genes per sample are normalized to the median, or an upper quantile, or positive control genes of a sample. Per gene normalization method normalizes each gene across conditions to a median in order to rescale the gene to a normalized range. In addition, probe intensities in microarrays are independent of each other, so the total intensities in each sample were not taken into account in the normalization of microarray data. In fact this is one of the differences from RNA-seq technologies, in which the reads of each transcript are affected by the total reads or read depth per sequencing sample.

Although the particular technical biases in microarray technologies are not present in RNA-seq approaches due to direct sequencing of the mRNA into short reads instead of its prospective hybridization to designed probe sets, there are other biological and technical variations where normalization is required. In RNA-seq, the expression level of each transcript is measured by the total number of mapped fragmented transcripts, which is expected to directly correlate with its abundance level. Therefore, the expected expression level of each transcript in its RNA-seq sample is limited by the sequencing depth or total number of reads, which is also pre-determined by the experimental design and budget before sequencing. For example, genes in the same sample with an average of 100X read depth will have about 3.3 fold higher read counts than one with an average of 30X read depth. Furthermore, transcript expression level is also dependent upon other transcripts within a sample [60], which means that given the fixed total read count, highly expressed transcripts will have a higher proportion of total reads [36, 61]. In addition, it is known that long transcripts will have more reads mapping to it than a shorter transcript of similar expression level within the same sample [62].

In the past several years, a number of normalization methods have been proposed to correct for library size bias as well as the length and GC-content bias. These methods include per-sample Total Counts (TC) [63], per-sample Upper Quartile (UQ) [58], per-sample Median (Med) [63], *DESeq* normalization [17], Trimmed Mean of M values (TMM) implemented in *edgeR* [36], Full Quantile (FQ) [64], Reads Per Kilobase per Million mapped reads (RPKM) [61], Fragments Per Kilobase per Million mapped fragments (FPKM) [65, 66], normalization by control genes [58, 67] and normalization by GC-content [68]. Most of these methods including TC, UQ, Med, DESeq and TMM use a common scaling factor per sample to normalize genes in order to correct library size. Among these, UQ, Med, FQ and control gene normalization were borrowed from microarray normalization. A few studies have reported preliminary results comparing statistical methods for normalization and DEGs in RNA-seq experiments using real or simulated data sets [58, 60, 63, 69-71], which show that the normalization methods have an important impact on the downstream analysis of DEGs. In fact, most normalization methods mainly address read depth by global normalization in addition to length and GC-content correction, but the length and GC-content bias associated with each gene is assumed to be constant and would be canceled out in the context of a differential analysis.

RNA-seq data are obtained from complex experiments with a variety of technical variations across different conditions. There are more variables that need to be adjusted for other than read depth [67]. The read counts of genes in each sample range from 0 or 1 for low abundant genes to high abundant genes with more than one million. In order to correct for variations of each gene across samples or conditions, we proposed a 2-step normalization procedure: correcting read depth through quantile normalization per sample followed by a per gene per 100 reads normalization across samples. This idea is adapted from the normalization of one-color cDNA microarray and RPKM in RNA-seq. The reads of each gene per sample are further scaled by the median per 100 reads across the conditions. Thus, the reads in each gene are put on the same scale which allows for an accurate comparison of gene expression across conditions. Since the exact test from *edgeR* [22] was used to test DEGs for comparing different normalization methods, the fraction of values after normalization by median of each gene across samples is corrected by multiplying 100 reads. This idea is borrowed from RPKM and FPKM normalization, which is rescaled by one million reads after scaling by the length of a transcript and total read counts of a sample. Therefore, the expected counts for lower expressed genes are increased, but decreased for highly expressed genes.

In this study, we proposed a new strategy for the normalization and named them as Med-pgQ2 and UQ-pgQ2, which are defined in two steps. In the first step, we performed quantile normalization per sample. The read counts of a gene in each sample were scaled by the medium reads of all genes defined as Med or by the upper quantile-reads of all genes in a sample defined as UQ. In the second step, we performed per gene normalization. The reads of each gene across samples were scaled by the medium reads of a gene defined as pgQ2. Then, we evaluated these methods (Med-pgQ2 and UQ-pgQ2) and other normalization methods (upper-quartile (UQ), full-quantile (FQ), *DESeq2* DESeq normalization, *edgeR* TMM (Trimmed Mean of M values) and *Cufflinks-Cuffdiff2*-FPKM). We used the exact test with a negative binomial distribution from *edgeR* for these normalization methods except the normalization methods from *DESeq2* and *Cufflinks-Cuffdiff2*. Our analysis used two public RNA-seq datasets sequenced by Illumina technologies, the Microarray Quality Control Project (MAQC) RNA-seq data [42] and simulated data based on breast cancer RNA-seq data with 24 normal and 25 early neoplasia (EN) samples [60, 72]. The results of this study suggest our alternative 2-step per gene normalization method to detect DEGs in RNA-seq data has higher overall power through greatly improving specificity and reducing false positive rates. This method may also be useful for combining RNA-seq data from different experiments using the same condition to increase the power for detection of DEGs via increasing sample size.

2. Materials and normalization methods

Normalization methods

The normalization methodologies adjust for variability at different levels, including within-sample, between-sample and other technical variability. Within-sample normalization enables the correction of expression level in each gene associated with other genes in the same sample. Since a long gene or transcript will have more reads mapping to it compared to a short gene or transcript of similar expression, length normalization is taken into consideration in some normalization methods. Currently, the most widely used within-sample and between sample normalization methods are Reads Per Kilobase of transcript per Million Mapped Reads (RPKM) [61] and Fragments Per Kilobase of transcript per Million mapped reads (FPKM) [65]. FPKM is used to count the reads of a fragment for paired-end RNA-seq data, which produces two mapped reads. Since length normalization is canceled out when identifying the DEGs across samples through a log ratio of a treatment versus a control usually called fold change (FC), many other normalization methodologies are used after that, which are described as follows.

As we know that the prominent variation of read counts for a gene between samples is due to the library size or the sequencing depth, the within-sample normalization of raw read counts is critical for the comparison of these gene expression measures across samples or experimental conditions. The simplest normalization method is called per-sample Total Counts (TC) normalization, which adjusts the raw read counts of each transcript by the total library size of each sample. However, studies in comparison of RNA-seq normalization methods show that more sophisticated normalization methods such as median, upper-quartile, TMM in *edgeR*, DESeq and full quartile normalization are much better than the TC normalization. One reason is that a small number of highly expressed genes can consume a significant amount of the total sequence [58]. To account for this feature, scaling factors are estimated from the data and used to achieve an inter-sample normalization [58, 63]. However, these methods have not considered between-sample normalization of each gene across conditions. The variation between genes within a sample and variation per gene across samples for these genes due to the systematical bias needs to be corrected in order to accurately identify the DEGs across conditions. In addition, Robinson *et al.* [22] demonstrated that the exact test for RNA-seq data is the best choice in terms of a small sample size achieving 5% nominal error rate compared to other test methods such as the Wald test , LR (Likelihood Ratio) test and asymptotic normal score test. Our proposed 2-step per gene normalization is able to center the reads of genes towards the median and minimize these variations.

In the following, we further define several statistical notations for characterizing different normalization techniques besides using some notations defined in the sample size calculations. For simplicity of the normalization methods, we only consider the gene g $(g = 1, \dots, G)$ in sample j $(j = 1, \dots, n)$ where G is the total genes and n is the total number of samples. Let X_{gj} be the number of observed reads mapped to a gene g for sample j, N_j be the total number of mapped reads for all genes in sample j, N be the total number of mapped reads across all samples, $\overline{N}$ be the mean of the reads across all samples and L_g be the length of the specific gene g.The above N_j, N and $\overline{N}$ can be expressed as:

$$N_j = \sum_{g=1}^{G} X_{gj},\ N = \sum_{j=1}^{n} N_j, \text{ and } \overline{N} = \frac{\left(\sum_{j=1}^{n} N_j\right)}{n}.$$

(1) Reads Per Kilobase of transcript per Million mapped reads (RPKM) [61]

Since a long gene will have more reads mapping to it compared to a short gene of similar expression, the length of the gene was considered in the RPKM normalization. Let X_{gj}^{RPKM} be the RPKM-normalized reads of gene g for sample j. The observed X_{gj} is scaled by both the total number of mapped reads (N_j) per million reads and the length of the transcript (L_g) per kilobase. Then X_{gj}^{RPKM} can be expressed as:

$$X_{gj}^{RPKM} = \frac{X_{gj}}{N_j \times L_g} \times 10^9 = \frac{X_{gj}}{\frac{N_j}{10^6} \times \frac{L_g}{10^3}} = \frac{Reads\ per\ transcript}{\frac{Total\ reads}{10^6} \times \frac{transcript\ length}{10^3}}. \ .$$

The RPKM normalization is actually a scaled normalization due to the different number of total reads as well as the different lengths of the gene g in sample j.

(2) Fragments Per Kilobase of transcript per Million mapped fragments (FPKM)

FPKM normalization is used for paired-end RNA-seq data and the relative expression of a transcript or gene is proportional to the number of cDNA fragments that originate from it. Let X_{gj}^{FP} be the FPKM-normalized reads of gene g for sample j. The observed X_{gj} is scaled by both the total number of mapped fragments (N_j) per million fragments and the length of the fragment of the gene or transcript (L_f) per kilobase. Then X_{gj}^{FPKM} can be expressed as:

$$X_{gj}^{FPKM} = \frac{X_{gj}}{N_j \times L_f} \times 10^9 = \frac{X_{gj}}{\frac{N_j}{10^6} \times \frac{L_f}{10^3}} = \frac{Fragments\ per\ transcript}{\frac{Total\ \#\ of\ fragments}{10^6} \times \frac{fragment\ length}{10^3}} \ .$$

The FPKM normalization is actually a scaled normalization due to the different number of total fragments as well as the different lengths of the gene g in sample j. The difference between RPKM and FPKM is that each mapped paired-end cDNA fragments will be counted as one in FPKM, but it will be counted as two reads in RPKM.

(3) Total count (TC) per sample

Let X_{gj}^{TC} be the TC-normalized reads of gene g in sample j. The observed X_{gj} is scaled by the total number of mapped reads $\left(N_j\right)$ per average of total reads across all the samples of the dataset$(\bar{N})$. Then X_{gj}^{TC} can be expressed as:

$$X_{gj}^{TC} = \frac{X_{gj}}{N_j} \times \bar{N} = \frac{X_{gj}}{N_j/\bar{N}}.$$

(4) Median (Med)

Let X_{gj}^{Med} be the median-normalized reads of gene g in sample j. The median normalization was based on all constitutive gene reads excluding any gene with zero counts for all samples. Then for each sample j, the normalization factor q_j^{50} is the median (50th percentile or 2nd quartile) of all the gene reads excluding zero-reads genes. The observed X_{gj} is scaled by q_j^{50} per average of median total reads across all the samples in the dataset $(\bar{N}_{med})$. Then, X_{gj}^{Med} can be expressed as:

$$X_{gj}^{Med} = \frac{X_{gj}}{q_j^{50}} \times \bar{N}_{med} = \frac{X_{gj}}{q_j^{50}/\bar{N}_{med}}. \qquad (45)$$

(5) Upper Quartile (UQ)

If there are many genes with very low read counts in a RNA-seq experiment, the upper-quartile normalization is preferred to the median normalization (50th percentile). Let X_{gj}^{UQ} be the upper-quartile-normalized reads of gene g in sample j. The upper-quartile normalization was based on all constitutive gene reads excluding any gene with zero reads for all samples. Then for each sample j, the normalization factor q_j^{75} is the upper-quartile (75th percentile) of all the gene reads excluding zero-reads genes. The observed X_{gj} is scaled by q_j^{75} per average of upper-quartile total reads across all samples in the dataset $\left(\bar{N}_{uq}\right)$. Then, X_{gj}^{UQ} can be expressed as:

$$X_{gj}^{UQ} = \frac{X_{gj}}{f_j^{UQ}} \times \bar{N}_{uq} = \frac{X_{gj}}{f_j^{UQ}/\bar{N}_{uq}}. \tag{46}$$

A study in evaluation of statistical methods for normalization in RNA-seq experiments [58] demonstrated that upper-quartile normalization reduced bias in the estimation of DEGs relative to qRT-PCR without noticeably increasing the level of variability as compared to total-count (TC) normalization.

(6) Trimmed Mean of M-value (TMM) [36]

Since we do not know the expression levels and true length of each transcript in RNA-seq data, RNA expression cannot be directly estimated from the raw read count. However, a relative gene expression level of two samples, i.e., a global fold change can be estimated. TMM is implemented in *edgeR* that assumes that a majority of the genes between samples are not differentially expressed. Let X_{gj}^{TM} be the TMM-normalized reads of gene g in sample j and S_j be the library size which is the unknown total RNA output of sample j.

In RNA-seq, a gene-wise log-fold-change of sample j relative to the reference sample r can be expressed as:

$$M_{gj} = log_2\left(\frac{X_{gj}/N_j}{X_{gr}/N_r}\right),$$

and absolute expression levels can be expressed as:

$$A_{gj} = \frac{1}{2} log_2^{(X_{gj}/N_j \times Y_{gr}/N_r)}, \text{ where } X_{gj} \neq 0.$$

A trimmed mean is the average after removing the upper and lower $x\%$ of the data. By default, the TMM procedure is doubly trimmed, where M_{gj} values are trimmed by 30% and A_{gj} values are trimmed by 5%. After trimming, a weighted mean of M_g is calculated and the normalization factor f_j^{TM} for sample j using reference sample r is expressed as:

$$f_j^{TMM} = \frac{\sum_{g \in G^*} w_{gj} \times M_{gj}}{\sum_{g \in G^*} w_{gj}}, \text{ where } M_{gj} = log_2\left(\frac{X_{gj}/N_j}{X_{gr}/N_r}\right),$$

and $w_{gj} = \frac{N_j - Y_{gj}}{N_j \times Y_{gj}} + \frac{N_r - X_{gr}}{N_r \times X_{gr}}$, where $X_{gj} > 0$ and $X_{gr} > 0$.

G^* represents the set of genes with valid M_{gj} and A_{gj} values, which is trimmed in advance of the calculation of M_{gj} and A_{gj} since log-fold-change cannot be calculated with zero transcript or gene reads in any selected samples $(X_{gj} = 0$ or $X_{gr} = 0)$. Therefore, G^* is not trimmed by the percentage above. Normalization factors across several samples can be calculated by selecting one sample as a reference and calculating the TMM factor for each non-reference sample. For example, for a two-sample comparison, only one relative scaling factor $\left(f_j^{TMM}\right)$ is required. It can be used to adjust both observed library sizes such as:

$$S_j^{TMM} = S_j \times \sqrt{f_j^{TMM}} \text{ and } S_r^{TMM} = S_r / \sqrt{f_j^{TMM}},$$

where $S_j = \sum_{g=1}^{G} u_{gj} \times L_g$ and $S_r = \sum_{g=1}^{G} u_{gr} \times L_g$.

The *calcNormFactors*() function in the *edgeR* Bioconductor package provides these scaling factors. The normalization factors are rescaled by the mean of the normalized library sizes as:

$$\bar{S}^{TMM} = \frac{\sum_{j=1}^{n} S_j^{TMM}}{n}.$$

X_{gj}^{TM} is obtained by using a re-scaled normalization factor to scale the raw read counts:

$$X_{gj}^{TMM} = \frac{X_{gj}}{f_j^{TMM} / \bar{S}^{TMM}}.$$

One notable difference with TMM normalization for RNA-seq is that the data themselves do not need to be modified and the estimated normalization factors are directly used in the statistical model to test for differentially expressed genes, while preserving the sampling properties of the data.

(7) DESeq normalization [17]

Like TMM, *DESeq* normalization is based on the assumption that most of the genes are not DE. Let X_{gj}^{DESeq} be the DESeq-normalized reads of gene g $(g = 1, \ldots, G)$ in sample j. A *DESeq* size factor f_j^{DESeq} given a sample i is calculated as the median of the ratios for the genes in each sample. The f_j^{DESeq} can be expressed as:

$$f_j^{DESeq} = \text{median}_g \frac{X_{gj}}{\left(\prod_{j'=1}^{n} X_{j'g}\right)^{1/n}},$$

where the denominator of this expression can be interpreted as a pseudo-reference sample obtained by taking the geometric mean across samples. Then X_{gj}^{DESeq} is obtained by scaling the raw reads X_{gj} by f_j^{DESeq}

$$Y_{gj}^{DESeq} = \frac{X_{gj}}{f_j^{DESeq}}.$$

(8) Full quantile normalization (FQ) [59, 73]

The full quantile normalization was originally used for the normalization of Affymetix GeneChip and one-color cDNA microarray [73]. The goal of quantile normalization is to make the distribution of probe intensities the same for arrays $j = 1, \ldots, n$, so that an *n*-dimensional quantile-quantile plot follows the *n*-dimensional identity line. The quantile normalization method of RNA-seq data consists of matching the distributions of gene or transcript reads across samples.

Let X_{gj}^{FQ} be the quantile-normalized reads of gene g in sample *j*. Let $q_g = \left(q_{1g}, \ldots, q_{ng}\right)$ be the vector of gth quantile for n arrays and $d = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ be the unit diagonal. The projection of q onto d is obtained by

$\text{proj}_d\, q_g = \left(\frac{1}{n}\sum_{j=1}^{n} q_{gj}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{gj}\right)$. This implies that each sample can be given the same distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset. The algorithm for normalizing a set of data vectors by giving them the same distribution is described as follows. 1) Let X_{gj} be the raw reads of genes and form a matrix of *X* with *G* x *n* dimension given *n* samples and *G* genes; 2) Sort each column of *X* to give X_{sort}; 3) Take the mean across rows of X_{sort} to obtain the mean reads per row and assign the mean to each row across samples to get X'_{sort}; 4) Getting the normalized $X^{QT}$by rearranging each column of X'_{sort} to have the same ordering as original *X*.

One possible problem with this method is that it forces the values of the quantiles to be equal. Therefore, it risks removing some of the signal in the tails. However, in practice, since gene expression measures are typically computed using the value of multiple transcripts, this may be not a problem for the gene in RNA-seq data.

(9) Per sample and per gene normalization

We propose a new strategy for normalization called "Per gene normalization after Median (Med) or 75% quantile (UQ) per sample normalization" and we name it Med-pgQ2 and UQ-pgQ2. The normalization procedures include three steps.

Step 1: Preprocess the raw count data

At first, genes with average raw counts less than the number of sample replicates are filtered out as a process of removing not expressed genes in both conditions. Then, the raw read counts are added by a small positive value (0.1 of one read) to generate a pseudo-count data that is used for the following normalization as well as the other normalization methods.

Step 2: Global per sample normalization

This step is to correct unequal library size using Med or UQ methods in equations (45) and (46).

Step3: Per gene normalization

This step is to compare each gene expression between conditions with detailed mathematical expressions as follows.

(a) **Med-pgQ2**: let X_{gj}^{Med} be the expression value for gene g in sample *j* and scaled by the median (Med) normalization method in equation (45); let $Q2_g^{Med}$ be the median of gene g across samples after Med normalization per sample. Thus, the new normalized counts $X_{gj}^{Med-pgQ2}$ per gene and per 100 reads can be expressed as:

$$X_{gj}^{Med-pgQ2} = \frac{X_{gj}^{Med}}{Q2_g^{Med}} \times 100. \quad (47)$$

With this method, the sample mean of gene g across samples will be changed by dividing $Q2_g^{Med}/100$ and the sample variance of gene g will be changed by dividing $\left(Q2_g^{Med}/100\right)^2$.

(b) **UQ-pgQ2**: let X_{gj}^{UQ} be the expression value for gene g in sample *j* and normalized by UQ (75%) normalization method in equation (46); let $Q2_g^{UQ}$ be the median of gene g across samples after UQ normalization. Thus, the new normalized counts $X_{gj}^{UQ-pgQ2}$ per gene and per 100 reads can be expressed as:

$$X_{gj}^{UQ-pgQ2} = \frac{X_{gj}^{UQ}}{Q2_g^{UQ}} \times 100. \quad (48)$$

With this method, the sample mean of gene g across samples will be changed by dividing $Q2_g^{UQ}/100$ and the sample variance of gene g will be changed by dividing $\left(Q2_g^{UQ}/100\right)^2$.

Statistical model and the exact test

A study by Robinson *et al.* [22] demonstrated that the exact test is the best method when the sample size is small, and results in achieving the nominal FDR compared to other methods such as the Wald test, the Likelihood Ratio test (LRT) and the asymptotic normal score test. In order to compare these normalization methods, we chose a negative binomial distribution to model and the exact test to identify DEGs for the majority of the methods using *edgeR*.

The negative binomial (NB) distribution. Briefly, $Y \sim NB(u, \phi)$ is a random variable to model the observed read counts in RNA-seq data, where Y has mean u and dispersion ϕ. Its probability mass function (pmf), the expected value and the variance of Y are correspondingly:

$$f_Y(y|u,\phi) = \mathrm{P}(Y = y) = \binom{y + \phi^{-1} - 1}{y} \left(\frac{1}{u\phi+1}\right)^{\phi^{-1}} \left(1 - \frac{1}{u\phi+1}\right)^{y},$$

$$E(Y) = u \text{ and } \mathrm{Var}(Y) = u + u^2\phi. \tag{5}$$

The above NB model utilizes the conventional parameterization called "NB2" [40]. The relationship of the parameters can be expressed as:

$$u = r\frac{1-p}{p} \text{ and } \phi = \frac{1}{r}.$$

Moreover, a negative binomial distribution can be derived from a Poisson-gamma mixture and hierarchy as:

$$Y|\lambda \sim Poisson(\lambda)\ and\ \ \lambda \sim gamma(\phi^{-1}, u\phi).$$

Then the marginal distribution of *Y* is a negative binomial and its pmf is expressed as the equation (12) [74].

Conditional dispersion estimation [22]: In this study, *edgeR* was used to evaluate the normalization methods. Since all libraries have the same library size after normalization (size factor equal to one), a CML (conditional maximum likelihood) is used in *edgeR* to estimate the dispersion parameter (ϕ_g) for a single gene *g* and sample *j* in *n* samples which is denoted as $Z = \sum_j^n Y_{gj} \sim NB\left(nu, \frac{\phi_g}{n}\right)$. It is expressed as:

$$l_{y|Z=z}(\phi_g) = \left[\sum_j^n \log\Gamma\left(y_{gj} + \phi_g^{-1}\right)\right] + \log\Gamma\left(n\phi_g^{-1}\right) - \log\Gamma\left(z + n\phi_g^{-1}\right) - n\log\Gamma\left(\phi_g^{-1}\right).$$

Exact test for a two-condition comparison in RNA-seq: The exact test for DEGs between two conditions is the best choice for RNA-seq [22]. Both *edgeR* [18] and *DESeq* implement similar exact test [17, 74]. Briefly, let Y_{gij} be the normalized counts of gene *g* in condition *i* = A and B, and replicate $j = 1, \ldots, n_i$. Then the assumptions concerning the distributions of Y_{gij} and $\sum_j^{n_i} Y_{gij}$ are expressed as:

$Y_{gij} \sim NB(u_{gi} \cdot s_{ij}, \frac{\phi_g}{s_{ij}}) \equiv Y_{gij} \sim NB(u_{gi}, \phi_g)$, where size factor $s_{ij} = 1$ for normalized counts,

$\sum_j^{n_i} Y_{gij} \sim NB(u_{gi} \cdot \sum_{j=1}^{n_i} s_{ij}, \frac{\phi_g}{\sum_{j=1}^{n_i} s_{ij}}) \equiv \sum_j^{n_i} Y_{gij} \sim NB\left(n_i \cdot u_{gi}, \frac{\phi_g}{n_i}\right), E\left(\sum_j^{n_i} Y_{gij}\right) = n_i \cdot$ u_{gi}, and $\mathrm{Var}\left(\sum_j^{n} Y_{gij}\right) = n_i \cdot u_{gi} + n_i \cdot u_{gi}^2 \cdot \phi_g$.

In order to identify DEGs between conditions A and B, we would like to test the null hypothesis H_0: $u_{gA} = u_{gB}$, where the test statistics are the total normalized counts in each condition, $Y_{gA} = \sum_j^n Y_{gAj}$, $Y_{gB} = \sum_j^n Y_{gBj}$, and total counts of two conditions $Y_{gS} = Y_{gA} + Y_{gB}$. The respective distributions of Y_{gA} and Y_{gB} are expressed as:

$Y_{gA} \sim NB\left(n_A \cdot u_{gA}, \frac{\phi_g}{n_A}\right)$, and $Y_{gB} \sim NB\left(n_B \cdot u_{gB}, \frac{\phi_g}{n_B}\right)$.

Since Y_{gA} and Y_{gB} are independent, the joint probability of $P(Y_{gA} = y_{gA}, Y_{gB} = y_{gB})$ under H_0 is $P(Y_{gA} = y_{gA}) \times P(Y_{gB} = y_{gB})$, In an exact test the p-value [17] is calculated by summation of the probability of a pair of *P(a, b)* that is less than or equal to the observed *P(y_{gA}, y_{gB})* given that the overall summation of *P(a, b)*. The pair of variables *a* and *b* are defined as $a = 0, \dots, Y_{gS}$ and $b = Y_{gS} - a$. Then the p-value for gene g is

$$p.value_g = P_{1g}/P_{2g},$$

where

$$P_{1g} = \sum_{\substack{a+b=Y_{gS} \\ P(a,b) \le P(y_{gA}, y_{gB})}} P(Y_{gA} = a) \times P(Y_{gB} = b),$$

and

$$P_{2g} = \sum\nolimits_{a+b=Y_{gS}} P(a,b).$$

The p-value is further adjusted by a false discovery rate correction. The dispersion parameter ϕ in equation (49) measures the extra variance of *Y* that a Poisson *(u)* distribution fails to describe. As ϕ goes to zero (ϕ→*0),* the variance of *Y* converges to *u* in probability and the distribution of $f(y)$ converges to the Poisson(u) distribution which was shown by Cameron and Trivedi [75].

Datasets

1). MAQC2 and MAQC3 datasets. MAQC2 contains two RNA-seq datasets from the Microarray Quality Control Project (MAQC) [76] with two types of biological samples: human brain reference RNA (hbr) and universal human reference RNA (uhr). The first dataset consisted of read length of 36bp and was downloaded from the NCBI sequence read archive (SRA) with ID SRX016359 (hbr) and SRX016367 (uhr) [58]. The second dataset (GEO series GSE24284) consisted of the 50bp hbr (sample ID: GSM597210) and uhr (sample ID: GSM597211) RNA samples [77].

GSE49712_HTSeq.txt.gz for MAQC3 raw read counts with five technical replicates in two biological conditions (UHR and HBR) was downloaded from GEO (GSE49712) [60]. Four replicate libraries for two conditions were prepared by one person and the remaining library was prepared by Illumina. A single HiSeq2000 instrument was used for sequencing all the samples.

2). TaqMan qRT-PCR data. PCR validation of the uhr sample from GSM12638 to GSM129641 and the hbr sample from GSM129642 to GSM129645 were downloaded from GEO (series GSE5350). These MAQC data (uhr and hbr) contain a total of 1044 genes assayed and validated using TaqMan qRT-PCR with 4 technical replicates [58, 77]. Thirty-seven of the 1,044 genes were marked with a Flag Detection "A" in all samples and were considered as true negative (TN) genes. These additional genes were not filtered out as in recent studies of the MAQC validation datasets [78] and 1028 of the

1044 genes have either a unique Ensembl gene Identifier (ID) or Entrez gene ID used for further analysis of the true positive and true negative genes following Bullard *et al.*'s study [58]. Briefly, a POLR2A-normalized cycle number for each gene and each condition is called ΔCt. The value x_{gik} of each gene g in replicate *i* and condition *k* is obtained via $log_2(\Delta\text{Ct})/\ log_2(e)$. The $\log_2$ fold change is defined as the mean difference of each gene between the hbr and uhr conditions ($\bar{x}_{g,hbr} - \bar{x}_{g,uhr}$), where the uhr is typically served as a reference. The genes with $|\log_2 FC| \geq 2$ were considered DEGs and the genes with $|\log_2 FC| < 0.2$ were considered as non-DEGs. Among the 1028 genes, 398 genes with 390 unique gene names fall into the true positive (TP) genes and 178 genes with 151 unique gene names fall into the true negative (TN) genes. The remaining set of genes lie in a region set to be indeterminate as far as DEG is concerned.

3). Two human breast cancer RNA-seq datasets. Dataset one is used for simulation containing twenty-four normal tissues and 25 early breast neoplasia (EN) on formalin-fixed paraffin-embedded tissue were sequenced at 3'-end enriched RNA-seq libraries [72].The mapped raw counts of 49 samples with an average of 7 million reads per sample were downloaded from NCBI GEO (series GSE47462). The dataset two contains 42 human estrogen receptor positive (ER+) and HER2 negative breast cancer primary tumors and 30 uninvolved breast tissues samples adjacent to ER+ primary tumors. The RNA-seq raw data files with a sequence read archive (SRA) were downloaded from NCBI GEO (GSE58135).

4). Simulated data. Simulated data was based on the human breast cancer RNA-seq dataset one with two conditions: 24 normal tissues and 25 early neoplasia tissues. The simulation model is similar to the one described in Dillies' study [63]. Let *G* be the total number of genes (G = 15,000), n = 20 be the total number of samples in two conditions (k = A, B), let y_{igk} be the count for gene g in sample i and condition k with a Poisson distribution: $y_{igk} \backsim Pois(\lambda_{gk})$. The parameter $\lambda_{gk}$is estimated from the mean reads per gene across samples from this human breast cancer RNA-seq dataset. Under this model, the null hypothesis H_0 $(\lambda_{gA} = \lambda_{gB})$ means the expression values of gene g between conditions A and B are not significantly different, and the alternative hypothesis H_1 $(\lambda_{gA} \neq \lambda_{gB})$ means the gene expression values are significantly different between the two conditions. Let p_0 and p_1 be the proportion of genes generated under H_0 and H_1 among the G genes, respectively. The data is simulated with 15,000 genes and p_1 is 10% corresponding to 1,500 genes. Under H_0, the parameter λ_{gA} in the gene g of condition A and the parameter λ_{gB} in the gene g of condition B were estimated from the breast cancer raw counts corresponding to the mean raw counts of each gene ($\lambda_{gA} = \lambda_{gB}$); while under H_1 the parameters λ_{gA} and λ_{gB} in the gene g and condition A and B were equal to $(1+\alpha)\lambda_{gA}$ for 750 downregulated genes and $(1+\alpha)\lambda_{gB}$ for 750 upregulated genes, respectively, where α is defined as 0.5 and 1. To assess the impact of non-equivalent library sizes, we multiplied y_{igk} by a size factor S_i per sample of the condition, which is equal to $|\varepsilon_i|$, where $\varepsilon_i \sim N(1,1)$. The number of simulation was chosen as 13 due to the small variation of the AUC values from all the normalization methods per simulation.

Sequence mapping and extraction of gene counts

The MAQC2 RNA-seq libraries with two technical replicates of each sample (uhr and hbr) and the human ER+ breast cancer dataset two were mapped to the human hg19 reference genome using *tophat2* (v2.0.13) with Bowtie version (2.2.3.0) and the parameter: ‘no-coverage-search’ [65, 79]. For the FPKM normalization method, the aligned RNA-seq reads were assembled according to the Homo_sapiens.GRCh37.74.gtf annotation file and normalized by FPKM using *Cufflinks-Cuffnorm* (v2.2.1). For the other normalization methods, the aligned RNA-seq reads were sorted by *samtools* (v0.1.19) and the read count matrix for each replicate of the condition was generated using HTSeq-scripts-count (version 2.7) and provided in S1_Datasets. In addition, for the human ER+ breast cancer dataset, the read counts from two human ER+ breast cancer samples and one control sample failed to be extracted using HTSeq-script-count. Therefore, only 40 ER+ breast cancer and 29 control samples were used for this study.

Software packages for detecting DEGs in normalization methods

The normalization methods and the software packages for detecting DEGs between conditions using MAQC datasets and the human ER+ breast cancer dataset are summarized in Table 13. Here, we give a brief description of the software packages used for the normalization and statistical tests in the present work. *edgeR* (v3.8.6) [18] was used to perform TMM normalization. It uses the empirical Bayes estimation and the exact test with a negative binomial distribution. For this study, edgeR was used to detect DEGs for all the seven normalization methods including TC, Med, UQ, FQ, TMM, Med-pgQ2 and UQ-pgQ2. *DESeq2* [80], a successor to the *DESeq* method [17], shows higher sensitivity and precision compared to DESeq package due to new features using shrinkage estimators for dispersion and fold changes. DESeq2 also offers a scaling size factor procedure as DESeq to perform normalization which is based on a median of ratio method. *Cufflinks-Cuffnorm* (v2.2.1) with a default parameter setting was used to perform FPKM normalization. Cufflinks-Cuffdiff2 was used to perform DEGs analysis at both the transcript and gene level using a beta negative binomial model and the t-test for the fragment counts [79]. In this study, we used the gene level results for the comparison with the other normalization methods. With the aid of edgeR, we set the normalization methods to “none” and selected the exact test with a tag-wise dispersion for each gene to perform DEGs analysis for the normalization methods: TC, Med, UQ, FQ, Med-pgQ2 and UQ-pgQ2. Since this study has been recently published, the normalized MAQC2 data from Med-pgQ2, UQ-pgQ2, DESeq and TMM-edgeR and DEGs analysis from these methods can be downloaded in Supporting Information S2-S5 Datasets [81]. Moreover, these normalization methods are written in R (v3.1.3) with the source codes publically available in S1 File (.R) [81].

Table 13: Summary of normalization methods and software packages on different datasets for DEGs analysis.

Normalization methods	Datasets	Statistical test	Software packages
TC	MAQC and simulated data	Exact test	edgeR(v3.8.6)
Med	MAQC and simulated data	Exact test	edgeR (v3.8.6)
UQ	MAQC and simulated data	Exact test	edgeR (v3.8.6)
FQ	MAQC and simulated data	Exact test	edgeR (v3.8.6)
TMM	MAQC and simulated data	Exact test	edgeR (v3.8.6)

Med-pgQ2	MAQC and simulated data	Exact test	edgeR (v3.8.6)
UQ-pgQ2	MAQC and simulated data	Exact test	edgeR (v3.8.6)
DESeq	MAQC and simulated data	Wald test	DESeq2 (v1.6.3)
FPKM	MAQC	t-test	Cufflinks-cuffdiff2 (v2.2.1)

The AUC, standard error and z-statistic test for MAQC data

The area under the ROC curve (AUC) was calculated using Algorithm 2 by Fawcett (2006) [82]. The estimated standard error (se) and a two-sample one-sided z-test were computed for each AUC value in MAQC data using Hanley J.A. *et al*. method (1982) [83]. Briefly, let *A* be the area under ROC curve; $\widehat{se}$ and *sd* be the estimated standard error and standard deviation, respectively; *na* and *nn* be the total number of true positive genes and false positive genes, respectively. Then, $\widehat{se} = \sqrt{d1/(na \times nn)}$, where $d1 = A \times (1-A) + (na-1) \times (Q1 - A^2) + (nn-1)(Q2 - A^2),\ Q1 = \frac{A}{2-A}, Q2 = 2 \times \frac{A^2}{1+A}$. The *Z* statistic was computed as: $z = \frac{A_1 - A_2}{\sqrt{\widehat{se}_1^2 + \widehat{se}_2^2}}$ and $p.value = 1 - Prob(Z < z)$. This p-value was used to compare the AUC values between two normalization methods.

The 95% confidence interval estimation of AUC for the simulated data

The 95% CI (confidence interval) for the simulated data was computed based on the normal approximation, which is defined as CI= $\bar{A} \mp 1.96 \times \frac{sd}{\sqrt{n}}$, where *n =13* is the number of simulations, $\bar{A}$ and *sd* are the mean and standard deviation of AUC from 13 simulations, respectively.

3. Results and Discussion

In this study, seven different normalization methods were compared to our proposed methods (Med-pgQ2 and UQ-pgQ2) via the qualitative characteristics of data distributions, intra-condition variation, ROC curve and AUC value as well as PPV, the actual FDR, sensitivity and specificity given the nominal FDR (≤ 0.05).

Qualitative characteristics of data distributions

In DEGs analysis, one important assumption of null hypothesis about normalized RNA-seq data is that the majority of genes are not differentially expressed between conditions. Therefore, the overall distributions across genes are expected to be similar. Boxplots of non-normalized $\log_2$ expression of raw read counts in Figure 5A shows larger distributional difference between the replicate libraries for MAQC2 data and normalization methods are needed to make the sample distributions more similar. Although all the normalization methods stabilized the distributions across two replicates for MAQC2 data, only our methods further can shrink the gene expression values towards the median per sample (Figure 5A).

It is important to compare the intra-condition variation among different normalization methods to prevent over correction. Figure 5B in MAQC2 data illustrates that little difference of the intra-condition variation is observed between our methods and others (DESeq, TMM, TC, Med and UQ), which indicates that scaling does not change the coefficient of variation. Moreover, we observed that FQ and FPKM methods greatly increased the intra-condition variation compared to the un-normalized data and other normalization methods (Figure 5B). This observation was also reported by Dillies' study in 2012.

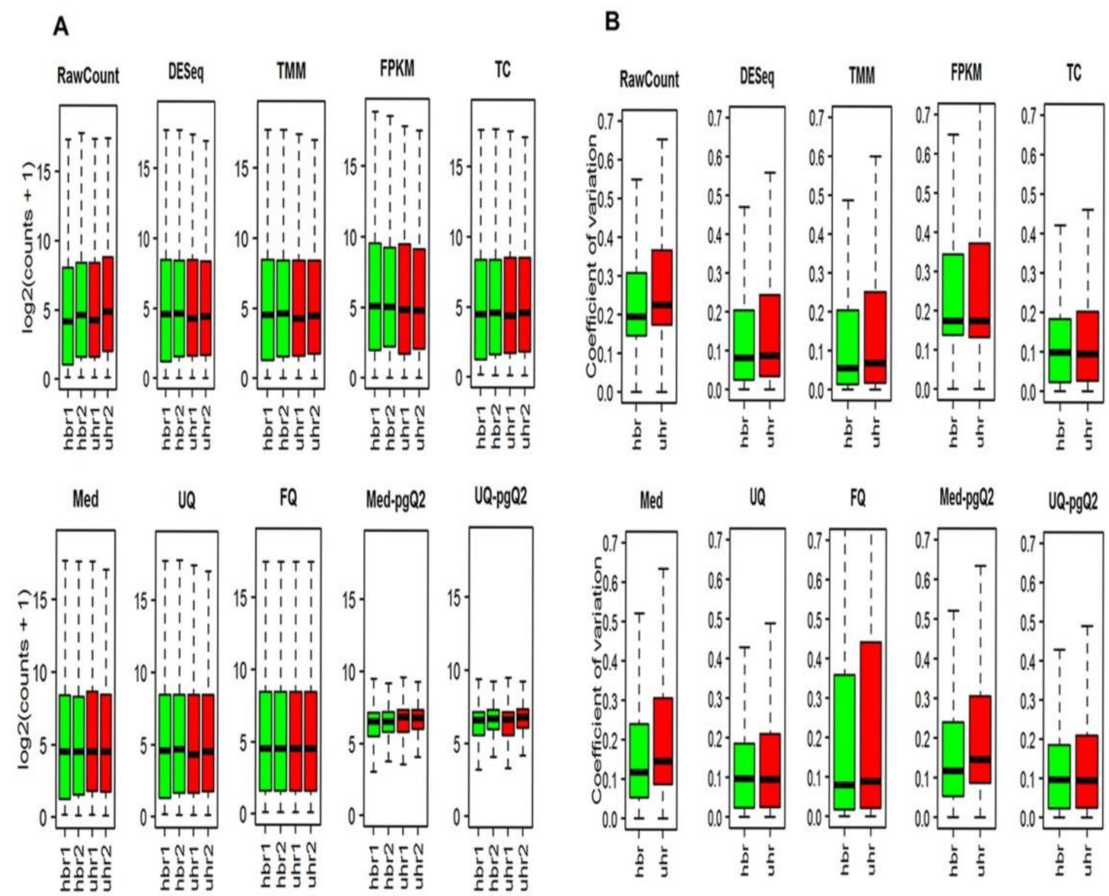

Figure 5: Comparison of nine normalization methods in boxplots. (A) Illustrated are boxplots of log2 (counts+1) for MAQC data with two replicates in two conditions (uhr and hbr). The samples in hbr and uhr conditions are in green and red, respectively. Med-pgQ2 and UQ-pgQ2 are our proposed methods. (B) Illustrated are boxplots of the intra-condition coefficient of variation (uhr and hbr), respectively.

We further analyzed the human ER+ breast cancer RNA-seq dataset with 40 ER+ breast cancer samples and 29 controls and MAQC3 with five technical replicates. For the human breast cancer datasets, similar patterns for most of the normalization methods from the boxplots (Figures 6 and 7) are observed compared to MAQC2 in Figure 5. However, the intra-condition variation of the median across replicates for all the methods (Figure 7) is close to 0.5, which is much higher than the value below 0.1 for all the methods obtained from the MAQC2 data (Figure 5B). This is expected because the breast cancer data contain biological replicates. We found that TC normalization failed in correcting the raw read counts for some of the replicates with a higher distributional difference within conditional replicates (Figure 6). The failed TC normalization was also observed by Dillies' study in 2012 using mouse miRNA-seq data. Furthermore, we also discovered that the inability of FQ normalization to minimize the intra-condition variation due to the small sample size from MAQC2 was diminished for the human ER+ breast cancer datasets with the sample size of 29 in control and 40 in ER+ breast cancer samples (Figure 7).

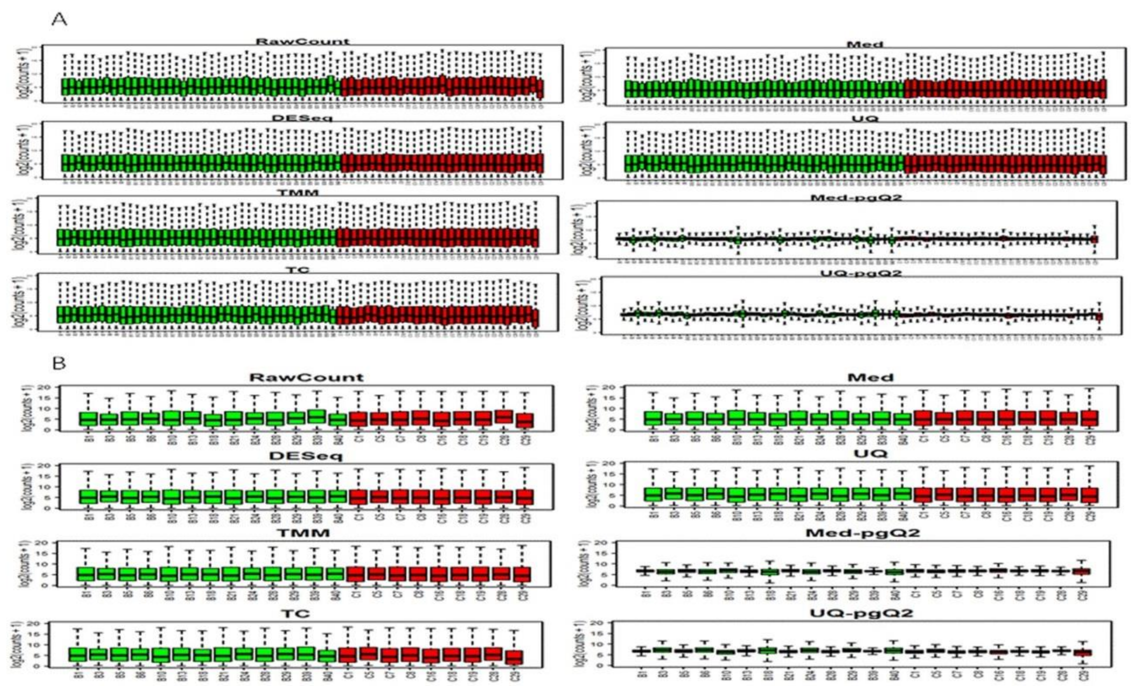

Figure 6: Data distribution from seven normalization methods using human ER+ breast cancer datasets.

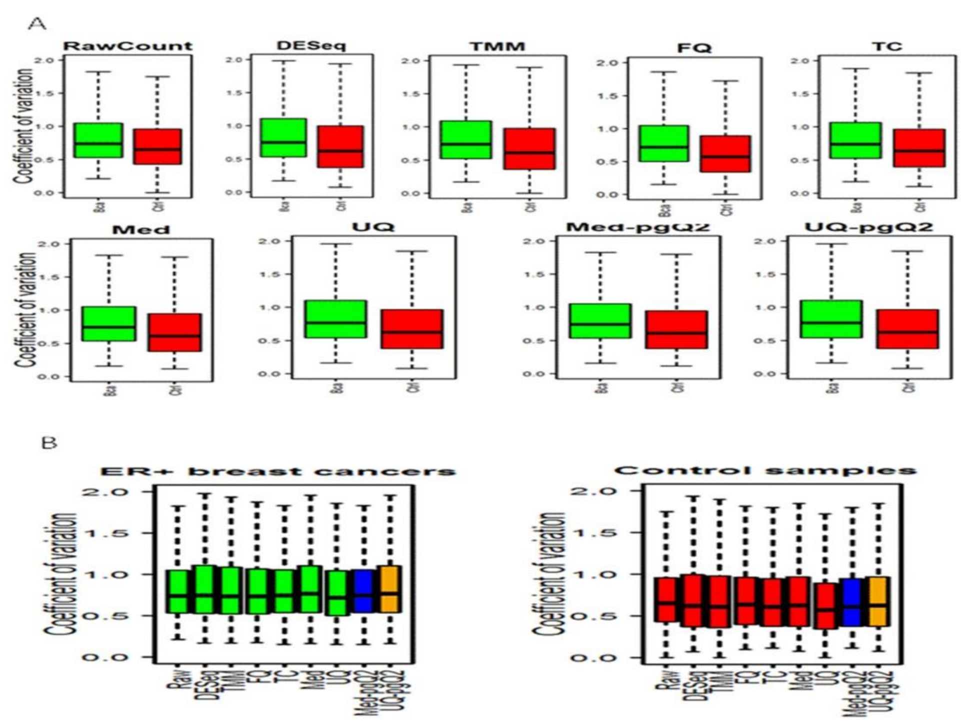

Figure 7: The intra-condition coefficient of variation using human ER+ breast cancer datasets.

For MAQC3 data, the boxplots (Figure 8) show that sample distributions normalized by all methods are very similar, which is expected due to technical replicates with very small variation. These data with less variation after scaling normalization suggest that a further per gene normalization may not show a great advantage other than shrinking the data toward the median across samples.

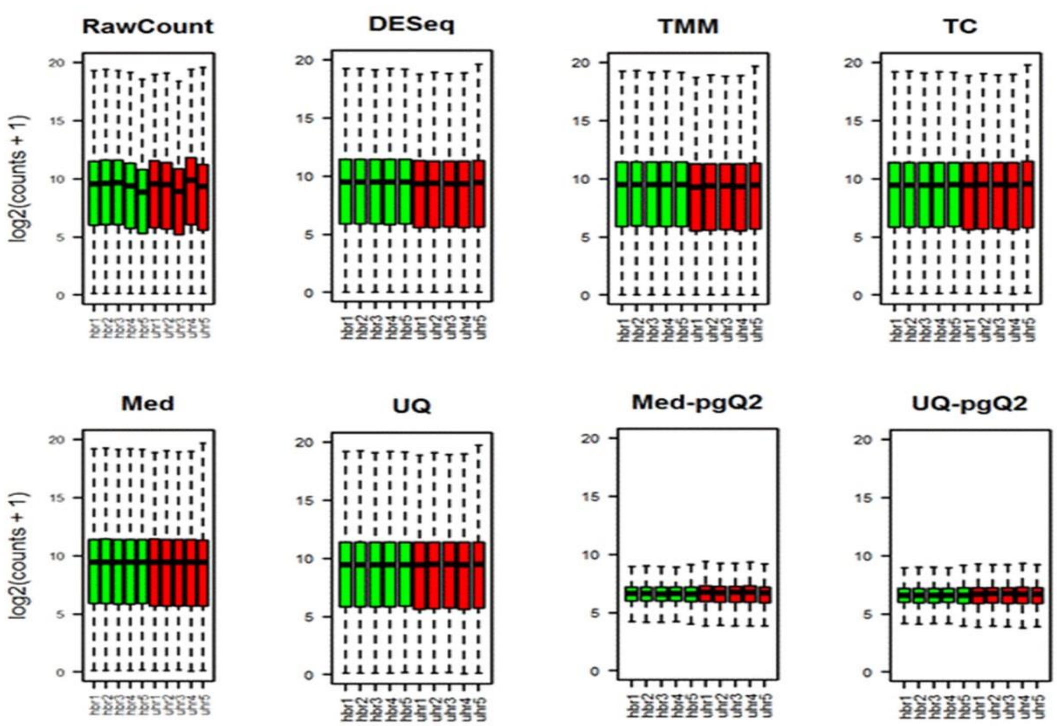

Figure 8: Data distribution from seven normalization methods using MAQC3 data.

RMSD between qRT-PCR and RNA-seq log_2 fold change computed by each method

To evaluate the accuracy of normalization methods, we used the MAQC2 and qRT-PCR data to calculate RMSD (root-mean-square-deviation) correlation between the $\log_2$ fold changes generated from statistical tests for each normalization method (**Table 13**) and the log2 fold changes from qRT-PCR. Figure 9 illustrates that almost all the normalization methods have good concordance to match the qRT-PCR data with RMSD accuracy less than 1.6 except *Cufflinks-Cuffdiff2* with a slightly higher RMSD value (1.77).

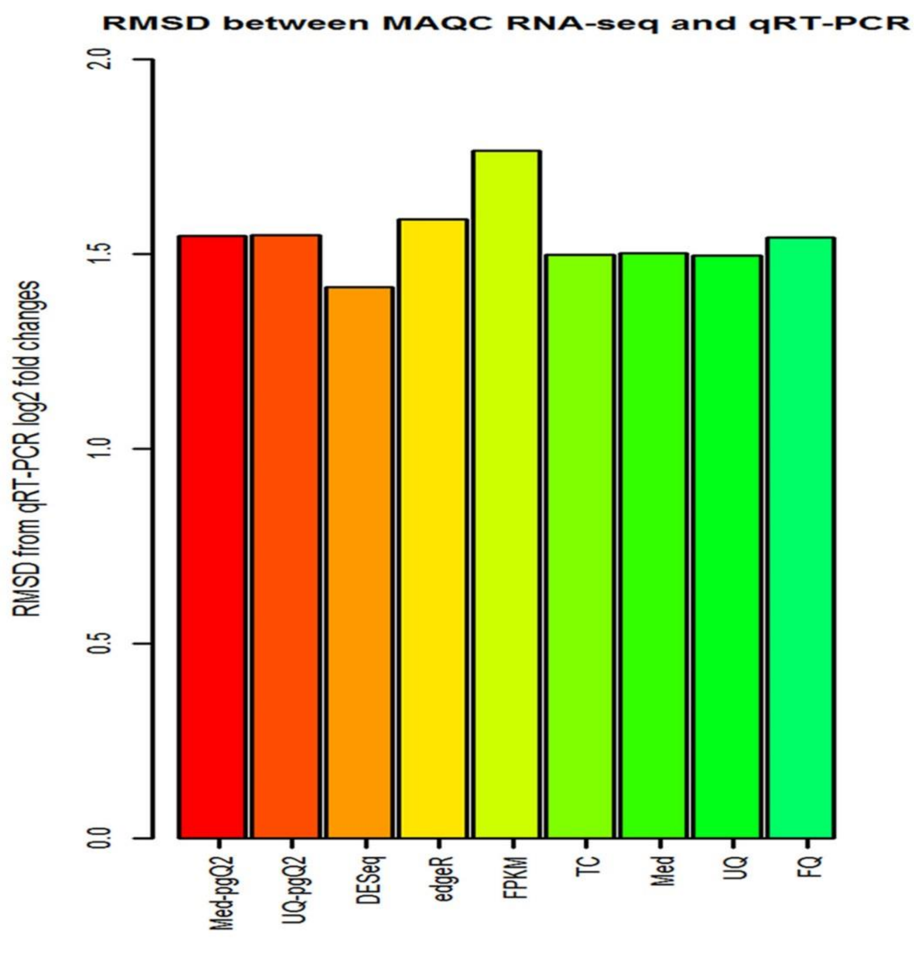

Figure 9: RMSD (root-mean-square deviation) between the $\log_2$ expression fold changes of MAQC2 and qRT-PCR. Illustrated is the RMSD between the $\log_2$ fold changes

computed from DEGs based on different methods and the values computed from qRT-PCR. FPKM (yellow) has the least similarity while DESeq normalization (brown) has the highest one.

Analysis of differentially expressed genes evaluated by ROC curves and AUC values

The ROC curve in Figure 10 is depicted by the relationship between the sensitivity and specificity rate based on MAQC2 data. The AUC value is calculated in the full range of false positive rate ($0 \leq FPR \leq 1$). Med-pgQ2 and UQ-pgQ2 achieve slightly higher AUC values compared to the others, which reflects the overall performance of detection of DEGs by achieving slightly higher sensitivity and specificity. With a false positive rate ≥ 0.10, the ROC cures reveal our methods perform slightly better. However, with a higher stringent false positive rate cutoff (< 0.10), the majority of the methods perform similarly. The quantile global normalization methods including TC, Med, UQ and FQ perform less favorable for this data. The standard error corresponding to the AUC value was also calculated using the equation from Hanley *et al.* in 1982 [83].

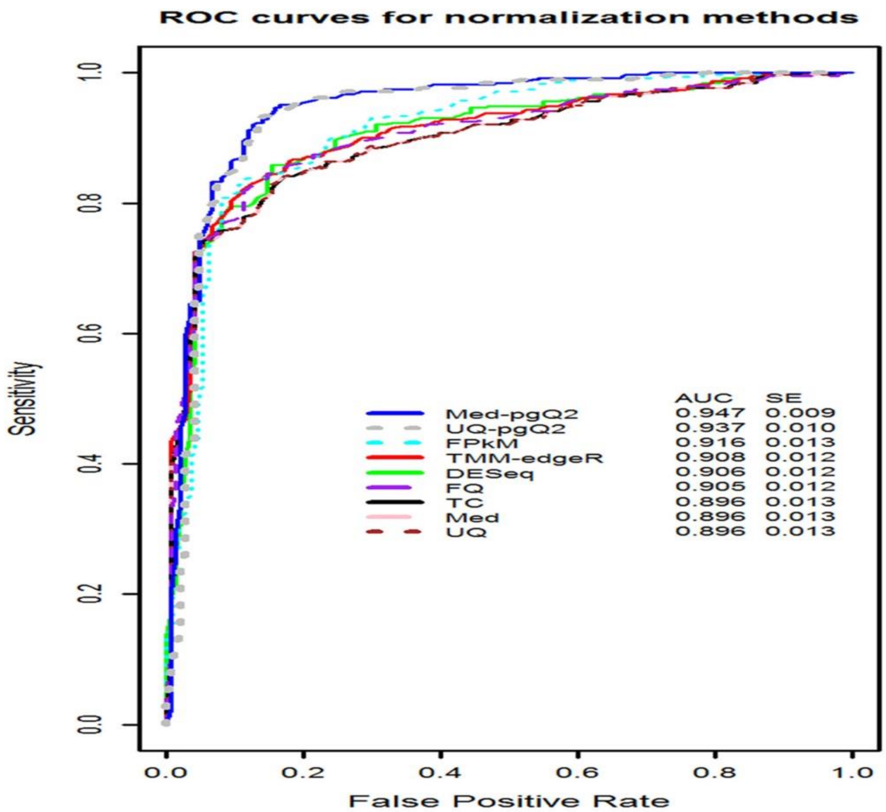

Figure 10: ROC curve and AUC values from MAQC2 data. The ROC curves and AUC values (inset) for evaluating the performance of the nine normalization methods were computed using MAQC2 with two conditions (uhr and hbr). Our proposed methods, Med-pgQ2 and UQ-pgQ2 (blue and grey, respectively) performed slightly better.

In addition, we further compared the AUC value from one of our methods (Med-pgQ2) to the others using a two-sample one-sided *z*-test. Table 14 lists the results of the p-values for each method. The results demonstrate statistically significant evidence that the AUC value in Med-pgQ2 is slightly larger than every other method except UQ-pgQ2.

Table 14: A one-sided of z-test on AUC values from Figure 10 comparing Med-pgQ2 to other methods.

	UQ-pgQ2	FPKM	TMM	DESeq	FQ	TC
z-statistics	0.7554	2.0082	2.0096	2.5826	2.6861	2.7517
p-value*	0.2250	0.0223	0.02224	0.0049	0.0036	0.0030

*p-values were computed using a one-sided of z-statistic test on the AUC values between Med-pgQ2 and one of the other methods listed in **Table 14**.

Analysis of PPV, actual FDR, sensitivity, specificity, and the number of true positive and false positive genes

In order to identify the major difference among all the normalization methods for detection of DEGs in MAQC2 and MAQC3 data, we calculated the number of true positive (TP) genes and false positive (FP) genes given the nominal FDR ≤ 0.05. We also calculated the positive predictive value (PPV), the actual false discovery rate (FDR), sensitivity and specificity for both datasets (Table 15). The results from MAQC2 data suggest that Med-pgQ2 and UQ-pgQ2 can achieve better specificity rate above 85% than other methods. While TMM-edgeR has the highest sensitivity rate (96.7%), its specificity rate (35%) is low. The performance of DESeq normalization with the sensitivity and specificity rate at 93.1% and 60.9% correspondingly are between our methods and TMM. The two proposed methods also achieve the lower actual FDR (< 0.1) compared to others. However, the results from MAQC3 with small variation in Table 3 show that all the methods achieve very high sensitivity rate above 98%, but the specificity for all the methods is lower than 42% and the actual FDR is higher than 0.15. The two new methods for these data perform slightly better in term of sensitivity, specificity and the actual FDR.

Table 15: Analysis of DEGs for MAQC2 and MAQC3 given a nominal FDR ≤ 0.05.

Datasets	Methods	# of TP genes	# of FP genes	Actual FDR	PPV	SR	SPR
MAQC2	DESeq	363	59	.140	.860	.931	.609
	TMM	377	97	.204	.797	.967	.358
	FQ	377	100	.210	.790	.967	.338
	TC, Med & UQ	376	101	.212	.788	.964	.331
	Med-pgQ2	362	22	.057	.942	.928	.854
	UQ-pgQ2	364	21	.055	.945	.933	.861
MAQC3	DESeq	385	105	.214	.786	.990	.271
	TMM	385	98	.203	.797	.990	.319
	TC, Med & UQ	384	99	.204	.795	.987	.313
	Med-pgQ2 & UQ-pgQ2	387	83	.177	.823	.995	.424

The number of true positive (TP) and the false positive (FP) genes, the actual false discovery rate (FDR), the positive predictive value (PPV), the sensitivity rate (SR) and specificity rate (SPR).

We further analyzed the DEGs detected only by the top performers such as DESeq, TMM and our methods using different quartile cutoff of mean expression of raw read counts from all genes given the nominal FDR ≤ 0.05. The results for the actual FDR, sensitivity and specificity are listed in Table 16. With the quantile cutoff at 75% by keeping the bottom reads in the analysis, the DESeq normalization has slightly better values in term of the actual FDR and specificity rate than other methods. TMM is least favorable in this case. With the quantile cutoff at 50%, DESeq outperforms others. With the quantile cutoff at 25%, TMM shows better performance than others and DESeq is relatively conserved. However, since there are a fewer genes listed as true positive and true negative genes at the quantile cutoff at 25% in MAQC2 data, this conclusion is not arbitrary. However, Table 16 suggests that our proposed methods (Med-pgQ2 and UQ-pgQ2) at the 100% quantile can achieve a sensitivity and specificity rate higher than 92% and 85% with the actual FDR less than 0.06, respectively. This study based on the MAQC2 data suggests our methods can improve specificity rate and the actual FDR for highly expressed genes. Based on the overall performance, it clearly indicates our methods might be the better choice for this kind of data.

Table 16: The actual FDR, sensitivity and specificity rate from MAQC2 data given a nominal FDR ≤ 0.05.

Expression quantile cutoff	**DESeq**			**TMM-edgeR**			**Med-pgQ2**			**UQ-pgQ2**		
	Actual FDR	SR	SPR	Actual FDR	SR	SPR	Actual FDR	SR	SPR	Actual FDR	SR	SPR
100%(total)	0.140	0.931	0.609	0.205	0.967	0.358	0.057	0.928	0.854	0.055	0.933	0.861
75%	0.069	0.861	0.806	0.147	0.931	0.516	0.084	0.877	0.758	0.077	0.898	0.774
50%	0.091	0.476	0.926	0.184	0.738	0.740	0.304	0.762	0.482	0.292	0.810	0.482
25%	0.000	0.000	1.000	0.333	0.333	0.917	0.667	0.667	0.333	0.692	0.667	0.250

The sensitivity rate (SR) and specificity rate (SPR) for DEGs analysis by the top methods at the different-quartile cutoffs.

To address the question of how gene-wise normalization methods (Med-pgQ and UQ-pgQ2) improve specificity while maintaining good detection power for highly expressed genes, we further analyzed gene-wise dispersion estimated after UQ and UQ-pgQ2 normalization with the aid of *edgeR* (Figure 11). Subsequently, gene-wise variance was estimated on the basis of the mean and estimated dispersion assuming a negative binomial distribution. We examined the coefficient of variation (CV) in two sets of genes based on a cutoff value of the mean read count (<100 vs. ≥100) from the UQ method. Genes with mean read count < 100 after UQ normalization were considered lowly expressed while the other genes were considered highly expressed. Figure 12 shows that the coefficient of variation for highly expressed genes after gene-wise normalization is increased via increasing the gene-wise dispersion and decreasing the per-gene mean read count compared to UQ normalization. This suggests that per gene normalization is more conservative for highly expressed genes, which at least partially explains our observation of improved specificity for these genes (Table 16). On the other hand, the coefficient of variation in lowly expressed genes after gene-wise normalization is slightly decreased compared to UQ normalization (Figure 11, bottom). This suggests that per gene normalization is less conservative for lowly expressed genes explaining our observation that our gene-wise normalization methods slightly improve sensitivity in this case (Table 16).

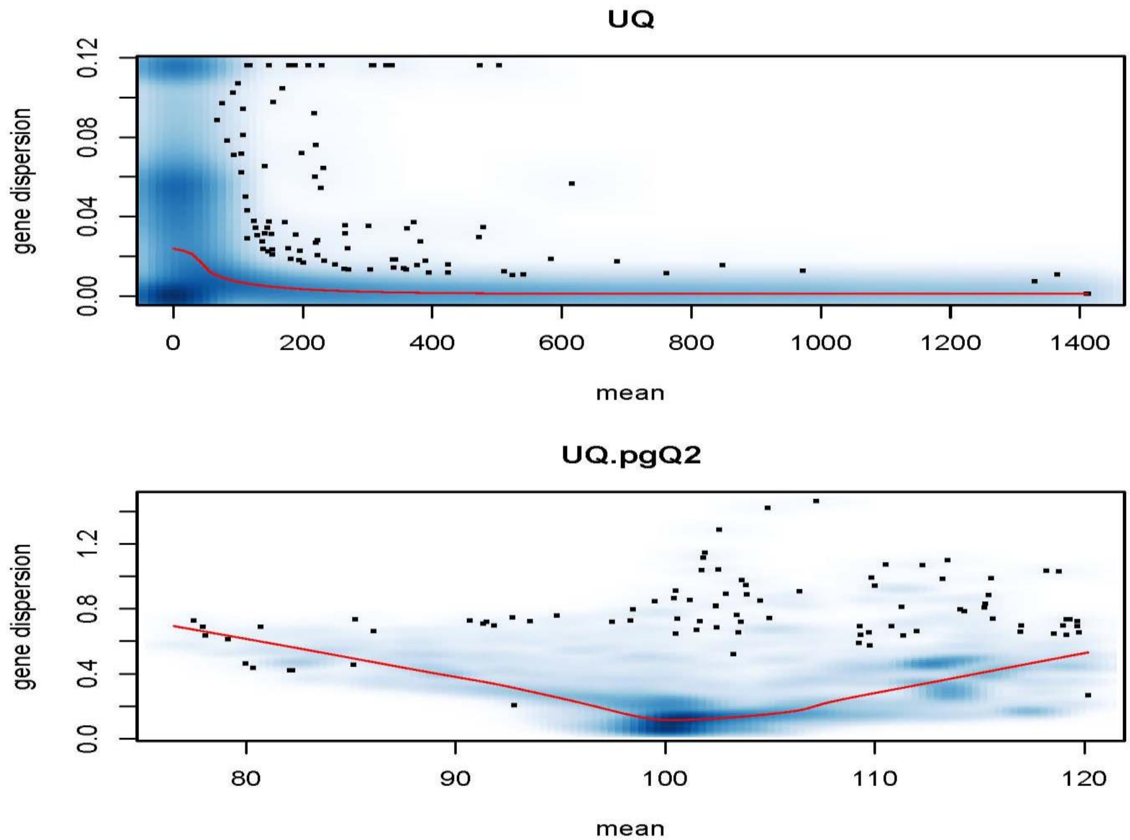

Figure 11: Mean vs. Dispersion after UQ and UQ-pgQ2 methods using MAQC2 data. Gene-wise dispersion was estimated after UQ and UQ-pgQ2 normalization with the aid of *edgeR*. The top graph displays mean versus gene dispersion for genes with a quantile cutoff value of mean read count after UQ normalization of ≤ 90%, while the bottom graph displays mean versus gene dispersion for genes with a quantile cutoff value of mean read count after UQ.pgQ2 normalization of ≤ 90%.

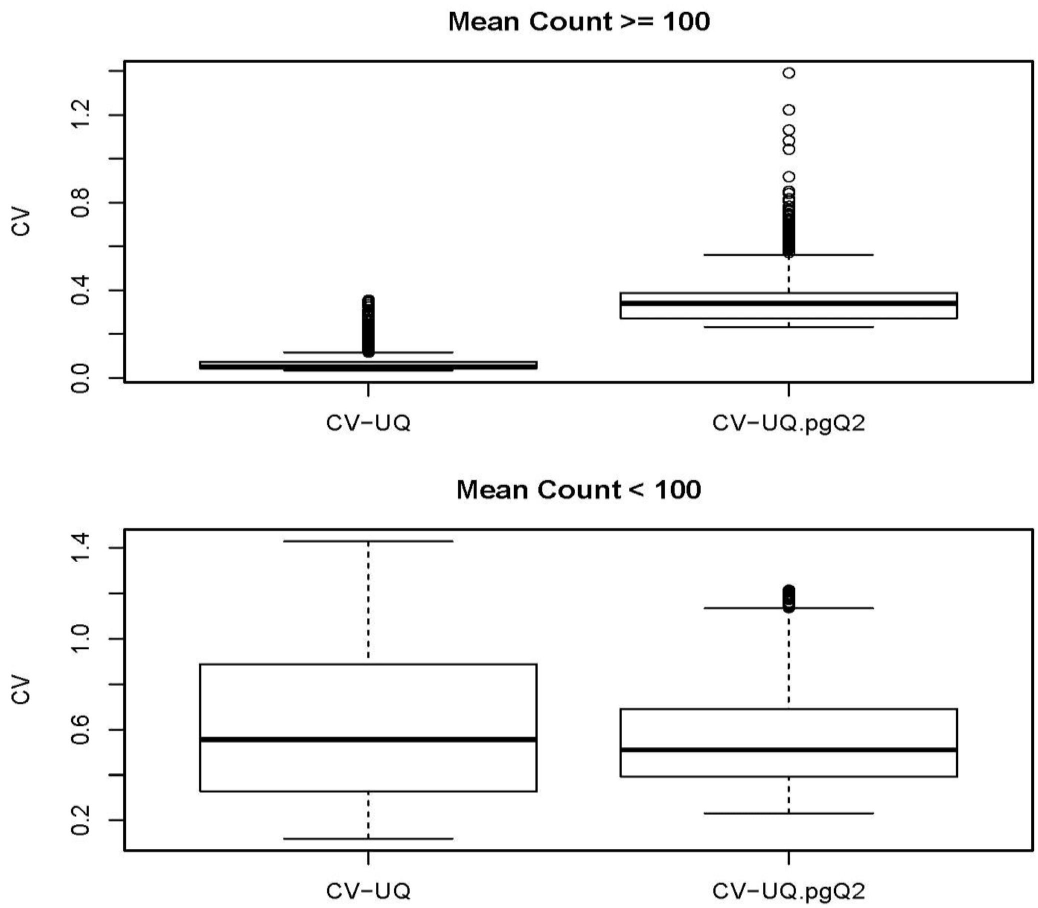

Figure 12: Coefficient of Variation (CV) after UQ and UQ-pgQ2 methods using MAQC2 data. The calculated coefficient of variation (CV) after UQ normalization and per-gene (UQ.pgQ2) normalization, based on the estimated dispersion parameter from *edgeR* and assuming a negative binomial distribution. The top graph displays the CV for genes with mean read count after UQ normalization of ≥ 100, while the bottom graph displays the CV for genes with mean read count <100.

Evaluation of normalization methods for detecting DEGs using different fold changes

The simulated data with 10 replicates and two conditions with different fold changes were used to compare our methods (Med-pgQ2 and UQ-pgQ2) based on the ROC curves. A total of 1,500 genes with a fold change (FC) of 1.5 and 2 are considered as true positive genes and the remaining genes (13,500) are considered as true negative genes. Fig 13A shows that the ROC curves for a FC of 1.5 in our methods have an average AUC value of 0.945 compared to others with the AUC value less than 0.924. Fig 13B shows that the ROC curves for a FC of 2 in our methods have the average AUC values greater than 0.980 compared to others with AUC values less than 0.969. However, the difference in the ROC curve and AUC values between our methods and others decreases as the fold change increases.

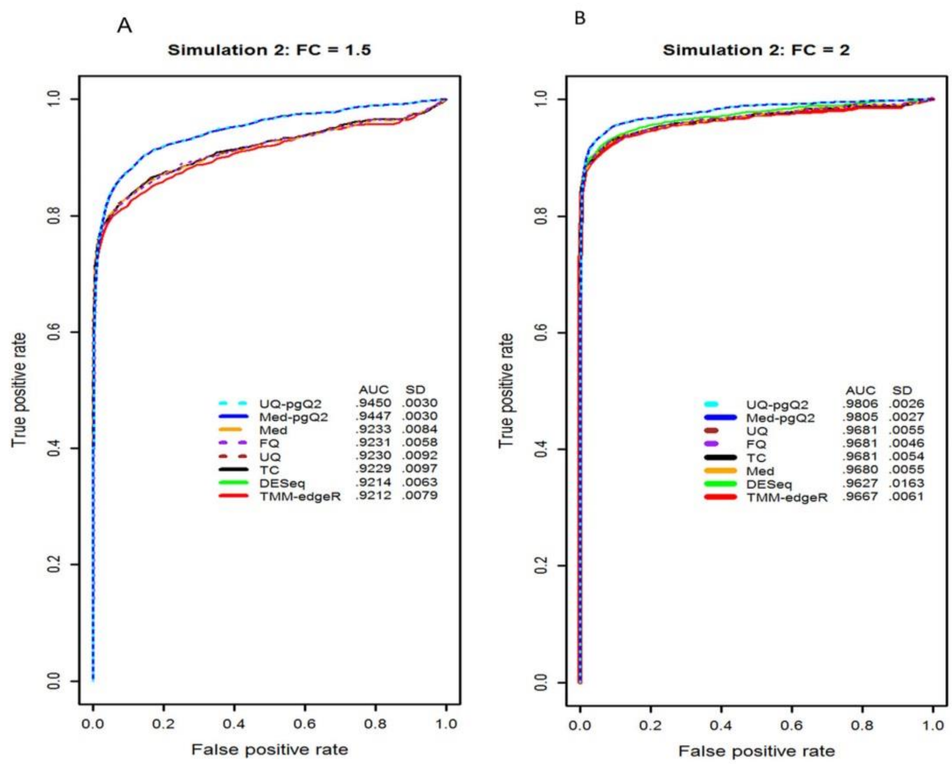

Figure 13: ROC curve and AUC values from the simulated data at a fold-change of 1.5 and 2. Illustrated are the ROC curves for detecting 1, 500 DEGs (750 up and 750 down-regulated) using a fold change = 1.5 (A) and a fold change =2 (B) with an unequal library size. Calculated AUC values are in the inset. The simulated data, containing a total of 15,000 genes in two conditions and 10 replicates per condition, was used for evaluating the performance of eight normalization methods. Our methods (UQ-pgQ2 and Med-pgQ2) are in cyan and blue, respectively.

Evaluation of normalization methods for detecting DEGs with biological replicates

We investigated the impact of biological replicates on the performance of normalization methods. We randomly sampled four and six replicates from each of 13 simulated datasets with 10 replicates used in Fig 13B, respectively. We sampled twice from one of 13 simulated data in Fig 13B yielding a total of 14 simulations. The mean AUC and standard deviation (SD) of each normalization method were calculated using 14 simulations instead of 13 simulations. The results from each simulation were consistent with a small standard deviation.

As expected, increasing the number of biological replicates yields a higher statistical power for detection of DEGs (Fig 14). Under the control of a very small false positive rate, the performance of all the methods (Med-pgQ2 and UQ-pgQ2) is similar. Fig 14 demonstrates that biological replicates are very important for RNA-seq data analysis in order to find true biological difference between conditions. Our normalization methods would be a good choice for achieving a slightly higher sensitivity rate at the false positive rate cutoff greater than 0.1. However, a closer examination for FPR cutoff less than 0.1 indicated that when the number of replicates is smaller (4 instead of 6), the other methods actually perform better than our proposed methods at a FPR cutoff less than 0.1 (Fig. 14A). This suggests that per gene normalization does not perform well for all circumstances. Therefore, caution is needed when choosing an optimal normalization method by taking into consideration the number of different replicates and their variation.

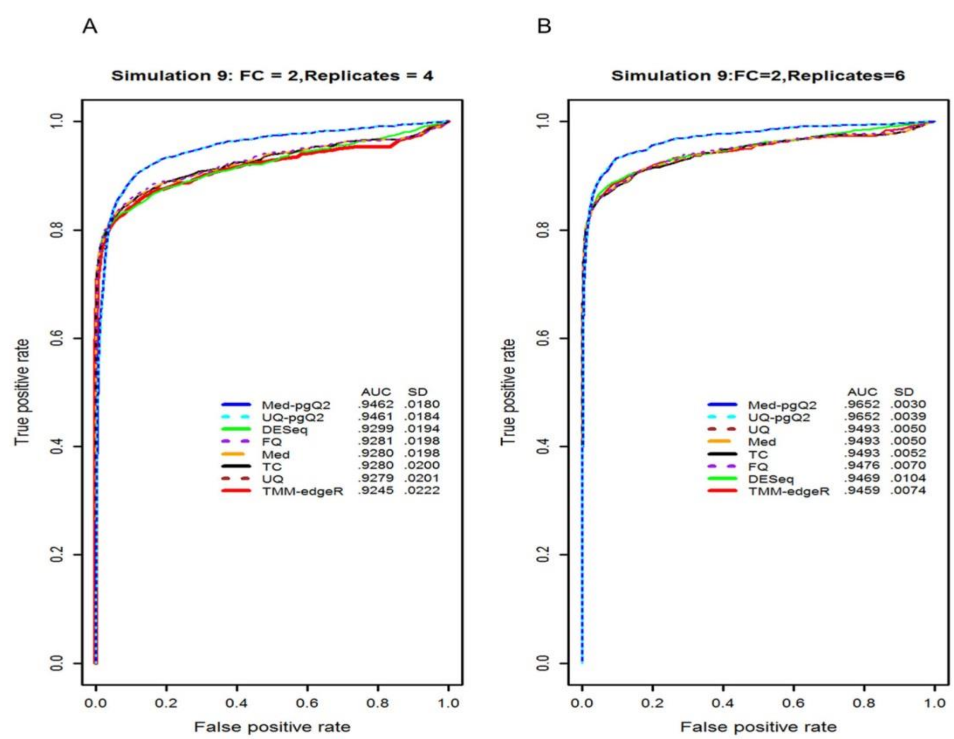

Figure 14: ROC curve and AUC values from the simulated data with 4 and 6 replicates in each condition. Illustrated are the ROC curves and AUC values (inset) in analyzing the impact of biological replicates on the performance of normalization methods. We used

the simulated data with four biological replicates (A) and six biological replicates (B), which contain 1,500 DEGs with 2 FC difference between two conditions. Our methods (UQ-pgQ2 and Med-pgQ2) are in cyan and blue, respectively.

Evaluation of Med-pgQ2 and UQ-pgQ2 methods for detecting DEGs in different multiplication factors (50, 100, 200, 500, 1000 and 1 million)

Like RPKM and FPKM, we chose to use the small multiplication factor of 100 for our proposed per gene and per 100 normalization for this study. We also chose different multiplication factors such as 50, 200, 500, 1000 and 1 million to perform per gene normalization. We performed DEGs analysis using Med-pgQ2 and UQ-pgQ2 with these multiplication factors. The comparison results based on DEGs analysis are shown in Table 17. We compared the impact of multiplication factors on PPV, the actual FDR, sensitivity and specificity. The values of PPV, the actual FDR, a sensitivity and specificity rate with multiplication factor ≥ 100 (Table 17) are more than 94%, less than 0.06, more than 92% and more than 85%, respectively. Little difference among them is observed except with the multiplication factor of 50 having a slightly higher sensitivity rate with a trade-off of a slightly higher actual FDR and a lower specificity rate. These results suggest that the choice of multiplication factors with a value greater than or equal to 100 has no difference on DEG analysis results.

Table 17: Evaluation of constant multiplication values for Med-pgQ2 and UQ-pgQ2 given the nominal FDR≤0.05. The number of true positive (TP) and false positive (FP) genes, positive predictive value (PPV), the actual false discovery rate (FDR), sensitivity and specificity for Med-pgQ2 and UQ-pgQ2 methods are computed using the MAQC2 data. These results of evaluation of the constant multiplication values (f=50, 100, 200, 500, 1000 and 1 million) are reported.

	Multiplication factor	# of TP genes	# of FP genes	PPV	Actual FDR	Sensitivity	Specificity
Med-pgQ2	50	365	23	0.941	0.059	0.936	0.848
	100	362	22	0.943	0.057	0.928	0.850
	200	362	21	0.945	0.055	0.928	0.861
	500	361	20	0.948	0.053	0.926	0.868
	1000	361	20	0.948	0.053	0.926	0.868
	10^6	360	20	0.947	0.053	0.923	0.868
UQ-pgQ2	50	364	24	0.938	0.062	0.933	0.841
	100	364	21	0.946	0.055	0.933	0.861
	200	363	21	0.945	0.055	0.931	0.861
	500	362	21	0.945	0.055	0.928	0.861
	1000	362	21	0.945	0.055	0.928	0.861
	10^6	361	20	0.948	0.053	0.928	0.868
DESeq	-	363	59	0.860	0.140	0.931	0.609
TMM-edgeR	-	377	97	0.796	0.205	0.964	0.358

4. Limitations

Our study has some limitations. First, the data normalized by med-pgQ2 and UQ-pgQ2 is restricted for DEGs analysis between groups and not for other purpose such as identifying highly or lowly expressed genes as well as comparing gene A to gene B expression levels within a sample due to the potential change of gene order in a sample after normalization. Second, a simulation data using a Poisson distribution based on real RNA-seq data with additional variation generated from a normal distribution was used for the DEG analysis. We do acknowledge that the lack of simulated data based on the NB distribution is a limitation to the study. However, inclusion of two real data sets (MAQC2 and MAQC3) offsets this limitation to an extent, and the combination of the simulated and real data provides fairly comprehensive and consistent answers. Finally, on one hand, the exact test was used to identify DEGs implemented by edgeR. Although it is recommended for DEG analysis of RNA-seq data in two groups with a small sample size, we think that evaluating the effect of normalization on more complicated study designs beyond two-group comparisons is a worthwhile and interesting endeavor, and we may consider this as potential future work. On the other hand, although a t-test is not commonly used for testing hypothesis in RNA-seq data, it is used for testing DEGs with small sample size in the cDNA Microarray data. Therefore, we need to mention here that a t-test is invariant to linear transformations and thus would be unaffected by the per-gene normalization outlined here.

5. Summary and Conclusion

Several studies have previously compared normalization methods (TC, Med, UQ, FQ, DESeq, TMM, FPKM and RPKM). TC, FPKM, RPKM and FQ are not suggested for use in DEG analysis due to multiple issues such as lowly expressed gene issue for TC, length correction bias for FPKM and RPKM, and potentially increasing the intra-condition variation by forcing all the samples to have identical distributions for FQ [58, 60, 62, 63]. One study has reported that UQ normalization failed to remove excessive variation from some of the samples [67]. DESeq and TMM-edgeR are in turn the only choices due to better performance compared to other existing methods. Although DESeq appears relatively conservative compared to TMM-edgeR method [71, 84], a high false-positive rate particularly for highly expressed genes for both methods has been observed by several studies [69, 78].

In this study, we compared two new normalization methods for RNA-seq data analysis (Med-pgQ2 and UQ-pgQ2) to the seven existing methods (DESeq, TMM-edgeR, FPKM-CuffDiff, TC, Med, UQ and FQ) based on DEG analysis. The purpose of using per-gene normalization approach is to remove technical variations using different chips and allow for comparison between conditions based on similar count levels [85, 86]. The results from this study demonstrate our proposed methods (Med-pgQ2 and UQ-pgQ2) can achieve a slightly higher value of AUC for both MAQC2 data and the simulated data at the false positive rate of 0.10, which reflects improving the overall performance with the detection power under the control of the low FDR compared to other normalization methods. More importantly, the results of DEG analysis from MAQC2 data with the different quantile cutoff values given a nominal FDR≤ 0.05, demonstrate our methods can decrease the false positive rate for highly expressed genes with high read counts giving the result of a specificity rate of greater than 85% without loss of a detection power (> 92%), while the other methods (i.e., DESeq and TMM-edgeR) have a specificity rate of less than 70%. Our methods may improve the sensitivity and detect more DEGs for lowly expressed genes with low read counts. However, given the improvement in the sensitivity for low read-count genes, there is a trade-off of a higher false positive rate in this case compared to DESeq and TMM-edgeR. Furthermore, the overall results from MAQC2 data also show the actual FDR from our methods is less than 0.06 while the actual FDR from DESeq, TMM-edgeR and others are larger than 0.10. This finding is consistent with the report by Kvam *et al*. in 2012. In their study they compared *DESeq*, *edgeR*, *baySeq* and *TSPM* (two-stage Poisson model) methods via a simulated data and reported the FDR in these methods are not controlled well and the actual FDR is larger than the observed FDR [69]. Moreover, we discovered DESeq and TMM have better overall performance than TC, Med, UQ and FQ, which is also consistent with previous studies. In addition, based on the quantile cutoff analysis of DEGs in MAQC2 data, we observed that DESeq is a good choice for moderately expressed genes at the quantile cutoff of 75%, but it is too conservative for lowly expressed genes at quantile cutoffs below 50%. However, TMM method seems to have better control of the false positive rate for the lowly expressed genes. In addition, the simulated study with four replicates shows that DESeq and TMM-edgeR methods perform better than our methods at the FPR cutoff less than 0.05. These new findings may give a better idea for the choice of different normalization methods.

There are several specific potential applications of our normalization methods worth mentioning. First, our methods may be useful for analyzing microRNA sequencing (miRNA-seq) data. Since miRNA expression is usually low compared to the mRNA with a ratio range 0.1~1.3% of total RNA in rat and mouse species, and 0.5~9.2% of total RNA in human samples, the data might be skewed to the low read counts. Therefore, per gene normalization may increase the sensitivity with a relative better specificity for detection of differentially expressed miRNAs [87, 88]. However, a comparison study of the performance for analyzing miRNA-seq between our methods and TMM-*edgeR* is needed to make definitive conclusions. Second, our methods are more universally applicable than using control-gene normalization in removing technical variations since it is hard to identify control genes such as housekeeping genes that remain at the same expression level regardless of the experimental conditions [63]. Third, given the importance of downstream analysis on RNA-seq data with a choice of normalization methods, our methods might be useful, particularly in light of emerging single-cell RNA-seq data and meta-analysis of RNA-seq data which have highly variable properties.

Finally, the simulated data results show that increasing the number of the biological replicates results in higher ROC curves and AUC values corresponding to higher detection power and lower false positive rate. However, due to the cost of RNA-seq data, the sample size of biological replicates was not considered by some of the earlier researchers using NGS technologies. One study by Hansen et al. in 2011 summarized a large number of published RNA-seq studies with a table showing that most of them had only one or a few biological replicates [11]. The thousands of DEGs identified from these RNA-seq data lack confidence and require further validation. Although laboratory qRT-PCR and Western blotting methods can be used to validate these identified DEGs, it is very tedious and almost impossible to validate several thousand DEGs. Our per gene normalization methods may be useful for combining the single or a few replicates of RNA-seq data from different experiments with the same conditions to increase the power for DEGs analysis.

Like many normalization and pre-processing procedures, our methods involve several choices of constants which we evaluated empirically. Primarily, in the 2nd step of our methods we chose to scale the median across samples to be per 100 reads instead of per kilobase or per million reads which was used by RPKM or FPKM. Our justification for this choice of a scaling constant in **Table 17** shows little difference of PPV, the actual FDR, specificity and sensitivity for multiplication factors ≥ 100 from DEGs analysis, and we picked the smallest scaling factor possible for which this was true. Secondly, a small positive value (such as 0.1 of one read) is added in all gene counts to avoid undefined fold changes in DEGs due to zero counts possible in one condition. This ensures no missing value for DEGs analysis and reduces the variability at low count values [89]. To study the robustness of results in the analysis of MAQC2 data, we considered different additive values (0.05, 0.1, 0.15, 0.2, 0.3, 0.4 and 0.5). The results in **Table 18** suggest that the FDR and sensitivity rate monotonically increased and the specificity rate monotonically decreased as increase in the additive values. Small positive values such as 0.05, 0.10 and 0.20 are recommended as FDR is reasonably maintained (less than 10%) with sensitivity and specificity rates of at least 80%. Furthermore, it is worth mentioning that preprocessing RNA-seq data such as prefiltering zero read counts across groups or adding a small positive number to all gene read counts is an option in RNA-seq data analysis. For example, the procedure to prefilter zero read counts may not avoid filtering out the lowly expressed genes which may be of interest by some researchers. Therefore, the choice of preprocessing the data will vary according to the experimental study.

Taken together, with the regards to all the discussed limitations, we think our proposed gene-wise normalization methods (Med-pgQ2 and UQ-pgQ2) might be a good choice for the skewed RNA-seq data with high variation via improving the false positive rate and maintaining a good detection power for DEGs analysis of RNA-seq data compared to the other normalization methods.

Table 18: Evaluation of the small positive values added in read counts for Med-pgQ2 and UQ-pgQ2 given the nominal FDR≤0.05. The number of true positive (TP) and false positive (FP) genes, positive predictive value (PPV), the actual false discovery rate (FDR), sensitivity and specificity for Med-pgQ2 and UQ-pgQ2 methods are computed from the MAQC2 data. The results for the choice of the small positive values added the read counts (f=0.05, 0.1, 0.15, 0.20, 0.30, 0.40 and 0.50).

	Added #	# of TP genes	# of FP genes	PPV	Actual FDR	Sensitivity	Specificity
Med-pgQ2	0.05	347	17	0.953	0.047	0.890	0.887
	0.10	**362**	**22**	**0.943**	**0.057**	**0.928**	**0.850**
	0.15	366	26	0.934	0.066	0.938	0.828
	0.20	369	30	0.925	0.075	0.946	0.801
	0.30	374	39	0.906	0.094	0.946	0.795
	0.40	377	45	0.893	0.101	0.962	0.722
	0.50	377	48	0.887	0.112	0.967	0.682
UQ-pgQ2	0.05	344	18	0.950	0.050	0.882	0.881
	0.10	**364**	**21**	**0.946**	**0.055**	**0.933**	**0.861**
	0.15	367	31	0.922	0.078	0.941	0.795
	0.20	372	35	0.914	0.086	0.954	0.768
	0.30	375	42	0.899	0.107	0.962	0.721
	0.40	376	50	0.882	0.117	0.964	0.669
	0.50	376	53	0.877	0.123	0.964	0.649
DESeq	-	**363**	**59**	**0.860**	**0.140**	**0.931**	**0.609**
TMM-edgeR	-	**377**	**97**	**0.796**	**0.205**	**0.964**	**0.358**

IDENTIFICATION OF THE OPTIMAL APPROACH FOR THE DIFFERENTIAL GENE EXPRESSION ANALYSIS IN HUMAN BREAST CANCER

1. Introduction

Breast cancer (BC) is the most commonly diagnosed cancer in women throughout the world [90-92], accounting for 23% of all female cancers [93-95]. BC is a growing health problem worldwide, increasing both in incidence [92] and resistance to treatment. Although significant progress has been made in the clinical treatment of BC, challenges persist because of its molecular heterogeneity, resistance to standard endocrine therapy and the risk of late recurrence. These challenges are driving intensive research efforts to identify new biomarkers of disease progression and signaling pathways amenable for potential chemotherapy.

Since BC is a heterogeneous disease, the identification of molecular pathologic markers, gene expression profiles and patterns of genomic alteration has become essential for predicting clinical outcomes and selecting appropriate therapy [96]. In particular, the presence of estrogen and progesterone receptors (ER and PR), and the human epidermal growth factor receptor 2 (HER2) have become standard biomarkers for defining BC subtypes which can be targeted by hormone modulation therapy. Approximately 75% of all BC are ER+ and of these, only half respond to anti-estrogen therapy [97, 98]. ER+ patients ultimately comprise the majority of deaths attributable to BC. Therefore, finding new targets for chemotherapy is an urgent need for preventive interventions [99-101]. Studies of ER+ BC reported that ER signaling engages in complex cross-talk with multiple signaling pathways via genomic and non-genomic regulation [97, 102]. ER+ BC has been characterized as having enhanced proliferation either by increasing cell division and/or decreasing apoptosis [102, 103]. On the other hand, tumors lacking ER and PR as well as HER2 (triple negative breast cancers, TNBC) are not amenable to these targeted therapies. Studies report that TNBC is more sensitive to chemotherapy than hormone positive BC [96, 104-106]. However, TNBC is associated with poorer survival than non-TNBC due to frequent relapse, and only about 31% of patients are completely responsive to chemotherapy [104, 105, 107].Therefore, a better understanding of the cell biology and molecular pathways underlying BC initiation and progression remains necessary for improving therapeutic options and clinical outcomes.

Mechanisms of oncogenesis involve the disruption of diverse biological functions and cellular pathways including cell cycle, proliferation, survival and apoptosis [94]. Recently, high throughput RNA sequencing (RNA-seq) has been increasingly used in clinical studies for defining gene expression changes [47, 49]. Indeed, RNA-seq-based gene expression profiling for the identification of global gene-expression patterns is a commonly used approach to integrate the multiple molecular events and mechanisms whereby cancer may develop [94, 108]. However, a standard procedure for differentially expressed gene (DEG) analysis remains undefined due to the multiple analytical steps required for RNA-seq workflow, including experiment design, library preparation and sequencing, mapping, normalization and statistical testing. Developing an appropriate approach to analyze RNA-seq data is challenging and critical. Studies comparing different normalization methods and statistical testing packages have shown that normalization methods and test statistics have a strong impact on DEG analysis [58, 60, 63, 69-71]. Using simulated data, it was found that the TC, UQ, Med, Quantile and RPKM normalization methods failed to control the false positive rate for genes with high read counts. In contrast, the DESeq (*DESeq2*) and TMM (*edgeR*) methods perform better overall than other methods in terms of detection power and control of false positives in data at a specified False Discovery Rate (FDR) [63]. However, using simulated data in these studies also report that an observed type I error rate is higher than the nominal FDR, leading to an inflated type I error rate. More recently, gene-wise normalization methods following per sample globally scaled normalization (UQ-pgQ2 and Med-pgQ2) were proposed by Li *et al*.[81]. A comparison of these methods with DESeq from *DESeq2* and TMM-edgeR normalizations using the benchmark Microarray Quality Control Project datasets (MAQC2) reported that Med-pgQ2 or UQ-pgQ2 perform slightly better for genes with high read counts via improving the specificity for skewed RNA-seq data given a FDR of 0.05. However, these methods also show a slightly higher FP rate for genes with a low read counts compared to *DESeq2* and *edgeR* [81].

In this study, we use a combinational and a robust approach to perform differential gene expression analysis using the ER+ with HER2- breast cancer (ER+HER2-BC) and TNBC datasets. First, we used *DESeq2* with DESeq normalization, *edgeR* with TMM normalization and UQ-pgQ2 normalization with a statistical test from *DESeq2*. We determined an optimal method via DEG analysis of within group comparisons given different absolute $\log_2$ Fold Change (|logFC|) cutoff values at a FDR ≤ 0.05. The DEGs identified from within group comparisons (e.g., 21 TNBC versus 21 TNBC) by a method are considered as false positive (FP) genes. We observe that the UQ-pgQ2 method with the aid of *DESeq2* performs best due to a fewer number of FP genes identified using the same |logFC| cutoff. *DESeq2* can detect slightly more FP genes than UQ-pgQ2, but *edgeR* performed the worst with many more FP genes for each within-group comparison. Thus, we selected UQ-pgQ2 and DESeq2 as the optimal methods for the DEGs analysis. To further decrease the FP rate in DEG analysis, we determined an optimal |logFC| cutoff at which the observed FP rate from within group comparisons is smaller than 0.01%. With this approach, the DEGs identified by both *DESeq2* and UQ-pgQ2 are considered as true DEGs. In addition, the DEGs identified either by *DESeq2* or UQpgQ2 at higher and more stringent |logFC| cutoff values are also considered as true DEGs. These higher |logFC| cutoff values are obtained when no DEGs or close to zero DEGs from the within-group comparisons are detected. Finally, three sets of DEGs from the combination analysis including two DEG sets uniquely identified in either of the TNBC or ER+HER-BC comparison groups and one common DEG set from both comparison groups are obtained. These DEGs are further analyzed for biological functions and pathways with the aid of the Ingenuity software.

In summary, our study reveals an important consideration to be made regard to the analysis of the BC datasets. One of issues is how to control the type I error rate with a small number set of high read counts of genes. Failing to consider this issue has a profound effect on the number of genes that are claimed as DEGs resulting in a misleading interpretation of biology. With the consideration of this important issue and using the knowledge base on the previous studies in a comparison of the existing methods for the analysis of DEGs, our analysis approach has identified UQ-pgQ2 normalization with the aid of DESeq2 is the best method for these BC datasets with high read depth given an absolute $\log_2$ Fold Change (|logFC|).We also found DESeq2 is a good method to analyze these BC datasets in terms of the number of DEGs detected with a slightly higher false positives. In our analysis scheme, we utilize UQ-pgQ2 and DESeq2 methods and robustly identify DEGs in TNBC vs. control and ER+HER2-BC vs. control. These DEGs deem to be truly and differentially expressed.

2. Materials and Methods

Normalization methods

In this study, we use three normalization methods including DESeq, TMM and UQ-pgQ2, and two software packages including *DESeq2* and *edgeR* for analyzing DEGs in human breast cancer datasets. DESeq and TMM normalization methods were implemented using *DESeq2* and *edgeR* packages. UQ-pgQ2 normalization was implemented using R [17, 18, 81].

Breast cancer datasets

The publicly available RNA-seq datasets contain forty-two Triple Negative Breast Cancer (TNBC) primary tumors; twenty-one uninvolved breast tissue samples that were adjacent to TNBC primary tumors (ctr1); forty-two Estrogen Receptor positive (ER+) and HER2 negative (HER2-) breast cancer (ER+HER2-BC) primary tumors and 30 uninvolved breast tissue sample that were adjacent to ER+HER2-BC primary tumors (ctr2). The RNA-seq raw data files with a sequence read archive (SRA) were downloaded from NCBI GEO (series ID GSE58135) [28].

Sequence mapping and extraction of gene counts

The raw SRA sequencing files containing 42 TNBC with 21 control samples (ctr1) and ER+HER2-BC with 30 controls samples (ctr2) were first converted .SRA files into .fastq files and then were mapped to the human hg19 reference genome using STAR (v2.5.3a). The mapped counts for 57,778 genes per sample were then extracted using HTSeq-scripts-count (version 2.7). After filtering the genes with zero counts across all the samples with four groups, 35,203 genes per sample were left for downstream analysis. Subsequently, the gene Symbol and Entrez gene ID with the type of protein are merged via the Ensembl gene ID.

Software packages used for normalization methods and testing DEGs

The normalization methods and the software packages with test statistics used for analyzing the human breast cancer datasets are summarized in Table 19. Here, we give a brief description of the software packages used for the normalization and statistical tests in the present work. *edgeR* (v3.8.6) [18] containing the TMM normalization method has been widely used for DEG analysis for RNA-seq data . *DESeq2* [80], a successor to the *DESeq* method [17], was used to implement DESeq normalization and a Wald's statistical test for detection of DEGs. For UQ-pgQ2 normalization, we choose *DESeq2* to perform a statistical test by setting the size factors of all samples as one and using a Wald's test to identify DEGs. Moreover, we used the gene level results for comparison with the other normalization methods.

Table 19: Summary of three normalization methods and software packages used for the analysis of breast cancer data.

Normalization method	Description of normalization	Distribution	Statistical test	Software packages
UQ-pgQ2	Per sample scaled by upper quantile and per gene by medium across samples	NB	Wald test	DESeq2 (v1.6.3)
DESeq	Per sample scaled by medium of ratio	NB	Wald test	DESeq2 (v1.6.3)
TMM	Per sample by Trimmed Mean M values	NB	Exact test	edgeR (v3.8.6)

Identification of an optimal normalization method via within-group comparisons

For determining an optimal normalization method with a control of false positives, we used UQ-pgQ2, *DESeq2* and *edgeR* to perform DEGs analysis of four within-group comparisons. The samples in each group were randomly selected and all the samples were divided into two groups. Thus, we have four comparisons: 21 TNBC vs. 21 TNBC, 11 ctr1 vs. 10 ctr1 (control for TNBC), 21 ER+HER-BC vs. 21 ER+HER2-BC, and 15 ctr2 vs. 15 ctr2 (control for ER+HER2-BC). We repeated the procedure 10 times via randomly selecting the samples into each group without replacement using R script. DEGs identified by these comparisons are considered as false positive (FP) genes.

We considered the method that could identify a much smaller number of FP genes using the same |logFC| cutoff value at a FDR ≤ 0.05 while maintaining a good detection power (> 10% total genes) as the optimal method. In order to minimize FP genes, a |logFC| cutoff for the identification of DEGs in the BC comparisons is determined when zero or fewer genes (less than 0.05%) identified both within BC and its control group.

For example, we arbitrarily chose the |logFC| cutoff of 2 for UQ-pgQ2 method when six and zero DEGs from a total of 35203 genes were identified in 21 TNBC vs. 21 TNBC and 11 ctr1 vs. 10 ctrl comparisons, respectively. Thus, the different optimal |logFC| cutoff values were selected for each method when comparing BC versus the control groups.

A robust approach to identify true DEGs for the comparisons of TNBC vs. crt1 and ER+HER2-BC vs. ctr2

1) After identifying UQ-pgQ2 and *DESeq2* as the optimal normalization methods, we performed the DEGs analysis in two scenarios: TNBC versus ctr1 and ER+HER2-BC versus ctr2 with the aid of *DESeq2*. The DEGs were determined in each comparison using an arbitrary |log FC| cutoff, where an observed FPR was $\leq 0.05\%$. 2) For each comparison, we assumed the DEGs identified from both normalizations as true DEGs. In addition, genes with a maximum (Max) |logFC| cutoff value identified either by *DESeq2* or UQ-pgQ2 were also assumed as true DEGs. This Max |logFC| cutoff value was determined when an actual or observed FPR was close to zero, which was based on the within-group analysis for the control group and the BC group 3) The true genes identified from TNBC and ER+HER-BC were further divided into three sets: the DEGs common in both TNBC and ER+HER2-BC; the DEGs only in TNBC or the DEGs only in ER+HER2-BC. These three sets of DEGs were further validated through the property of the subtypes of BC.

Biological function and pathway analysis

I used Ingenuity Pathway Analysis (IPA) software to validate the DEGs identified using my analysis scheme. On the other hand, I used IPA to identify cancer-related and statistically significant biological functions and canonical signaling pathways for DEGs identified in the comparison of TNBC and ER+ groups (QIAGEN, version 3355999, USA).

3. Results and discussion

1). DEGs identified from within-group comparisons to determine the optimal normalization method and statistical packages.

DEGs identified from the within-group comparisons of BC data using UQ-pgQ2, *DESeq2* and *edgeR* are listed in Table 20. The results with a different |logFC| cutoff show that UQ-pgQ2 is more conserved than other methods in most cases.

When comparing 21 TNBC versus 21 TNBC, the number of FP genes with a |logFC| $\leq$ 1.5 cutoff detected by UQ-pgQ2, *DESeq2* and *edgeR* was 4 (± 4), 43(± 34) and 527(± 125), respectively, with a FPR in a total of 35203 genes of 0.01%, 0.12% and 1.50%. Similarly, given a |logFC| cutoff of 2, the number of FP genes within the TNBC comparison using UQ-pgQ2, *DESeq2* and *edgeR* was 0, 10(± 9) and 455(± 97), respectively, with a FPR of 0, 0.03% and 1.29% correspondingly. The number of FP genes within the ctr1 comparison using UQ-pgQ2, *DESeq2* and *edgeR* was 0, 0 and 6(± 12), respectively. In this case, UQ-pgQ2 and *DESeq2* achieved a FPR below 0.05% both within group comparisons from TNBC and control groups. Thus, the |logFC| of 2 was chosen as an optimum cutoff value for the downstream analysis of DEGs. With a |logFC| cutoff of 2.5, the number of FP genes within the TNBC comparison detected by

UQ-pgQ2, *DESeq2* and *edgeR* was 0, 1(± 1) and 372(± 74), respectively, with a FPR of 0, 0.003% and 1.06%. In this case, UQ-pgQ and *DESeq2* achieved a FPR below 0.01%. Given a |logFC| ≥ 3, *edgeR* had an observed FPR of 0.84% within the TNBC comparison.

For the 21 ER+HBR-BC versus 21 ER+HBR-BC comparison, the number of FP genes with a |logFC| ≤ 1.5 cutoff using UQ.pgQ2, *DESeq2* and *edgeR* was 1 (± 1), 6(± 3) and 292(± 67), respectively, with a FPR in a total of 35203 genes of 0.003%, 0.02% and 0.83% correspondingly; and the number of FP genes in the within ctr1 comparison from UQ-pgQ2, *DESeq2* and *edgeR* was 1(± 3), 14(± 16) and 771(± 184), respectively, with a FPR of 0.003%, 0.04% and 2.19%. In this case, UQ-pgQ2 and *DESeq2* achieved an observed FPR below 0.05% both within the group comparisons (ER+HER2-BC and control). Thus, the |logFC| of 1.5 was chosen as an optimum cutoff value for the DEGs analysis. Moreover, given a |logFC| cutoff of 2, the number of FP genes within the ER+HER2-BC comparison from UQ.pgQ2, *DESeq2* and *edgeR* was 0, 1(± 1) and 259(± 54) with an FPR of 0, 0.003% and 0.74% , respectively; and the number of FP genes within the ctr1 comparison from UQ.pgQ2, *DESeq2* and *edgeR* was 0, 5(± 6) and 686(± 137). In this case, UQ-pgQ2 and DESeq2 achieved a FPR below 0.01% both within the ER+HER2-BC and control group comparison. With a higher |logFC| cutoff of 2.5, the number of FPR genes within the ER+HER2-BC comparison detected by UQ.pgQ2, *DESeq2* and *edgeR* was 0, 0 and 216(± 39), respectively; and the number of FP genes within the ctr1 comparison from UQ-pgQ2, *DESeq2* and *edgeR* was 0, 1(± 2) and 590(± 94), respectively. In this case, UQ-pgQ and *DESeq2* achieved the observed

FPR below 0.01%, but *edgeR* has an observed FPR above 0.7% both within the ER+HER2-BC and control group comparisons.

In summary, an approach via the within group analysis to identify FP genes can help us to achieve several goals. First, among the choice of three methods, we considered UQ-pgQ2 and *DESeq2* to be better for DEGs analysis of the BC datasets while UQ-pgQ2 performs the best in term of controlling false positives. Second, the results (Table 21) helped us to choose an optimal |logFC| by taking a consideration of a FPR and a high detection power with a reasonable number of DEGs. Finally, the results (Table 20) indicate that UQ-pgQ2 was slightly better conserved than *DESeq2* while *edgeR* is too liberal. This observation is consistent with the previous studies comparing normalization and statistical test methods for the DEGs analysis of RNA-seq data.

Table 20: DEGs analysis within-group and between groups using UQ-pgQ2, *DESeq2* and *edgeR*. The DEGs are determined using a different |log FC| cutoff at a FDR≤0.05.

	Comparison groups	UQ-pgQ2	DESeq2	edgeR
|Log(FC)|≥1.5	21TNBC vs. 21 TNBC	4±4	43 ±34	527±125
	11ctr1 vs. 10 ctr1	1±2	0	6±14
	21 ER+HER2-BC vs. 21 ER+HER2-BC	1±1	6± 3	292±67
	15 ctr2 vs. 15 ctr2	1±3	14±16	771±184
	42 TNBC vs. 21 ctr1	7,474	8,969	9,585
	42 ER+HER2-BC vs. 30 ctr2	4,999	6,308	7,448
|Log(FC)| ≥2	21TNBC vs. 21 TNBC	0	10±9	455± 97
	11ctr1 vs. 10 ctr1	0	0	6±12
	21 ER+HER2-BC vs. 21 ER+HER2-BC	0	1±1	259 ±54
	15 ctr2 vs. 15 ctr2	0	5±6	686±137
	42 TNBC vs. 21 ctr1	3,706	5,201	5,854
	42 ER+HER2-BC vs. 30 ctr2	2,169	3,176	4,161
|Log(FC)| ≥2.5	21TNBC vs. 21 TNBC	0	1±1	372± 74
	11ctr1 vs. 10 ctr1	0	0	5±10
	21 ER+HER2-BC vs. 21 ER+HER2-BC	0	0	216±39
	15 ctr2 vs. 15 ctr2	0	1±2	590±94

	42 TNBC vs. 21 ctr1	1,701	2,888	3,586
	42 ER+HER2-BC vs. 30 ctr2	869	1,499	2,326
|Log(FC)| ≥3	21TNBC vs. 21 TNBC	0	0	296± 54
	11ctr1 vs. 10 ctr1	0	0	4±7
	21 ER+HER2-BC vs. 21 ER+HER2-BC	0	0	175±33
	15 ctr2 vs. 15 ctr2	0	0	502±65
	42 TNBC vs. 21 ctr1	767	1815	2,290
	42 ER+HER2-BC vs. 30 ctr2	323	689	1,356

Table 21: The number of DEGs from two methods based on the total of 35,203 genes with an observed FPR given a |logFC| cutoff in parenthesis.

	Normalization	**DEGs**	**FPR (|logFC|)**	**a |logFC| given FPR≈ 0**
TNBC	UQpgQ2	3,706	≈0 (≥ 2)	≥ 2
	DESeq2	5,201	≈0.03% (≥ 2)	≥ 2.5
	Common DEGs	3,610	-	-
ER+HER2-BC	UQpgQ2	4,999	≈0.003% (≥ 1.5)	≥ 2
	DESeq2	6,308	≈0.04% (≥ 1.5)	≥ 2
	Common DEGs	4,776	-	-

FPR: an observed false positive rate

2). DEGs identified between-group comparison in human TNBC and ER+HER2BC using UQ-pgQ2, *DESeq2* and *edgeR* based on 35203 genes

Gene expression profiles in two comparisons (42 human TNBC versus 21 ctr1, and 42 human ER+HER2-BC versus 30 ctr2) were analyzed using three methods (UQ-pgQ2, *DESeq2* and *edgeR*), where *DESeq2* was used for preforming statistical test for UQ-pgQ2 normalization method.

DEGs were identified giving a nominal FDR ≤ 0.05 and an optimal |logFC| cutoff value (Tables 20-21). The results in Table 20 show that *edgeR* has a higher detection power while having a tradeoff of a higher FPR given the same |logFC| cutoff according to the within-group analysis. Although *DESeq2* performs better in terms of FPR when compared to *edgeR*, previous studies report the actual type I error in *DESeq2* is higher than the nominal FDR, particularly in high read counts of genes. With the aid of *DESeq2,* UQ-pgQ2 has a much lower FPR while have a tradeoff of a fewer number of DEGs detected. In order to maximize the detection power and minimize the type I error, we select UQ-pgQ2 and *DESeq2* to identify the DEGs given a 0.05 FDR and an optimal |logFC| cutoff (Table 21).

The results in Table 21 show that by using UQ-pgQ2 method, 3706 DEGs in TNBC and 4999 DEGs in ER+HER2-BC were detected given an optimal |logFC| cutoff with an observed FPR below 0.002%. Similarly, using *DESeq2*, 5,201 DEGs in TNBC and 6308 DEGs in ER+HER2-BC were detected given the same |logFC| cutoff as UQ-pgQ2 with an observed FPR below 0.03%.

3). DEGs identified in human TNBC and ER+HER2-BC from UQ-pgQ2 and DESeq2 based on 17584 protein coding genes.

Gene expression profiling is commonly used to identify disease biomarkers and biological functions. In RNA-seq data analysis, we noted that the identified DEGs contain a mixture of mRNA, miRNA, rRNAs and other non-coding RNAs that are present in the total RNA per sample and not completely eliminated during the library preparations. These noncoding RNAs, especially about 10% of high abundant rRNAs with high read counts remain in each RNA-seq sample while using the ploy-A depletion method in the library preparations. In this study, we focus on the 17,584 protein coding genes out of the 35,203 total genes. The results in Table 22 show that using UQ-pgQ2 method, the number of DEGs detected for TNBC at a |logFC| $\leq$ 2 and ER+HER2-BC at a |logFC|$\leq$ 1.5 was 1,584 and 2,303, respectively; using *DESeq2* method with the same cutoff values, the number of DEGs detected for TNBC and ER+HER2-BC was 1,913 and 2,649, respectively. The number of DEGs common in both analytical methods for TNBC and ER+HER2-BC was 1546 and 2212, respectively. In addition, Table 22 also shows the number of up and down-regulated DEGs per comparison.

Table 22: DEGs identified using *DESeq2* and UQ-pgQ2 based on 17524 protein coding genes with a nominal FDR $\leq$ 0.05 and an arbitrary |logFC| cutoff.

Data	Normalization	DEGs	Up	Down
TNBC	UQ.pgQ2	1,584	949	635
	DESeq2	1,913	1099	814
	Common DEGs	1,546	915	631
ER+HER2-BC	UQ.pgQ2	2,303	1,161	1,142
	DESeq2	2,649	1,195	1,454
	Common DEGs	2,212	1,074	1,138

4). A robust method to identify the true DEGs (protein coding genes) based on 17,854 protein coding genes using UQ.pgQ2 and DESeq2 methods

Based on the DEG analysis of the 17,584 protein coding genes, we noted that the number of DEGs identified by the UQ-pgQ2 and *DESeq2* methods varied for the two comparisons (42 TNBC versus 21 control and 42 ER+HER2-BC versus 30 control). The previous studies observed that *DESeq2* and *edgeR* are less conserved for the high read count genes using MAQC2 data. Therefore, in order to minimize the number of false positives, we used a combinational approach to identify the true DEGs. The results are listed in Table 21. For the TNBC comparison, we first identified the common DEGs using both UQ-pgQ2 and *DESeq2* resulting in 1546 DEGs. Similarly, for the ER+HER2-BC comparison, 2212 DEGs were identified in common. We assumed the DEGs that were not in common, but identified by either *DESeg2* or UQ-pgQ2 with an observed FPR close to zero given a Max |logFC| cutoff (Table 21), were also considered as true DEGs. With this approach, in the TNBC comparison, 109 DEGs from *DESeq2* at a |logFC| ≥2.5, and 38 DEGs from UQ-pgQ2 at a |logFC| ≥2, were considered as true DEGs, and adding them to the common DEGs set results in 1,693 true DEGs. For the ER+HER2-BC comparison, with a |logFC| ≥2, 84 DEGs from *DESeq2* and 3 DEGs from UQ-pgQ2 are considered as the true DEGs, and adding them to the common DEG set results in 2299 true DEGs.

The heatmaps (Figure 15) based on the DESeq-normalized gene expression levels were constructed using a hierarchical clustering from Partek software (Partek Genomics Suite 6.6). In this figure, we conventionally chose the up-regulated genes in red and down-regulated genes in green. Figure 15A illustrates the gene expression level of the 1,693 DEGs for the 42 TNBC versus 21 control samples. Figure 15B illustrates the gene expression level of the 2,299 DEGs for the 42 ER+HER2-BC versus 30 control samples.

Finally, we examined the number of common and unique DEGs between 1,693 DEGs in TNBC and 2,299 DEGs in ER+HER2-BC.We found the 896 DEGs that were common in both TNBC and ER+HER2-BC, while the 797 DEGs and the 1403 DEGs were uniquely identified in TNBC and ER+HER2-BC, respectively. These common and unique DEGs were further used for the analysis of the cancer-related biological functions and pathways with the aid of Ingenuity.

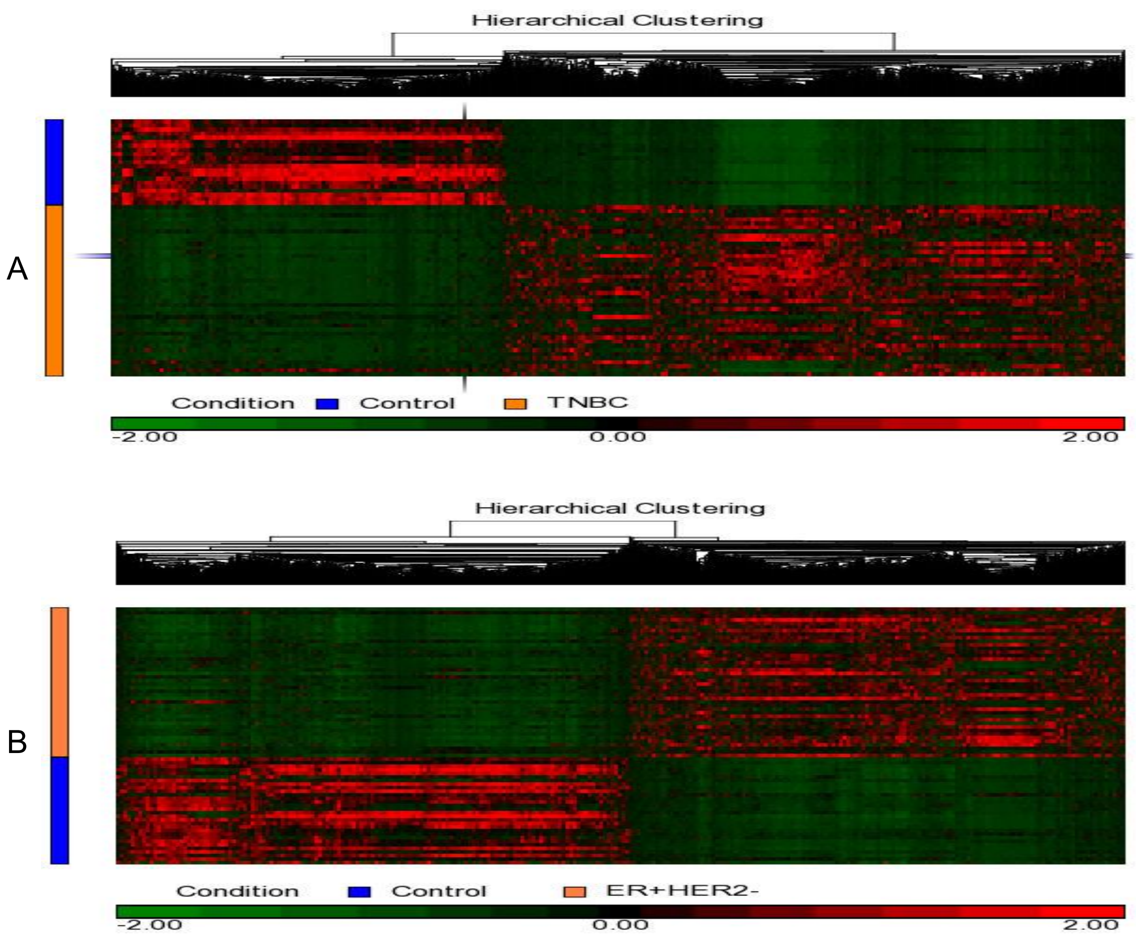

Figure 15: Heatmaps based on the DESeq-normalized gene expression levels are constructed. The up-regulated genes are in red and the down-regulated genes are in green. A. The1693 DEGs identified from TNBC. B. The 2299 DEGs identified from ER+HER2-BC.

5). Identification of biomarker genes based on the presence or absence ER, HR and HER2 to partially validate the DEGs analysis

We identified biomarker genes based upon the presence or absence of the molecular receptors. The results are listed in Table 23. For the TNBC comparison, we found that ER (ESR1 and ESR2), PR (PGR) and HER2 (EGFR) were significantly down-regulated using both UQ.pgQ2 and *DESeq2* methods as expected. For the ER+HER2-BC comparison, we found that ER1 (ESR1) was significantly up-regulated with a FC greater than 1.8 and ER2 (ESR2) is significantly down-regulated. PR (PGR) expression level in ER+HER2-BC was not significantly different from the control groups. However, HER2 (EGFR) was significantly down-regulated using both methods as expected. Taken together, the expected results via the molecular markers can partially validate the true DEGs using an integrated approach.

Table 23: Biomarker for TNBC and ER+HER2-BC.

Comparison	Symbol	LogFC (UQ.pgQ2, DESeq2)	FDR (UQ.pgQ2, DESeq2)
TNBC	ERS1	(-2.95,-3.25)	≤ 0.001
	ERS2	(-0.94,-1.01)	≤ 0.002
	PGR	(-3.12, -3.56)	≤ 0.001
	EGFR (HER2)	(-1.51, -1.65)	≤ 0.01
ER+HER2-BC	ERS1	(0.92,0.84)	≤ 0.005
	ERS2	(-1.84, -2.01)	≤ 0.001
	PGR	(0.23, -0.36)	≥ 0.59
	EGFR (HER2)	(-3.23, -3.43)	≤ 0.001

6). Top cancer-related biological functions identified via Ingenuity software

Ingenuity Pathway Analysis (IPA), which uses information obtained from high impact journals, is a widely used tool for the partial validation, but mainly used in identification of diseases and biological functions. The three sets of DEGs containing the 797 protein coding genes uniquely identified in TNBC, the 1403 protein coding genes uniquely identified in ER+HER2-BC and the 896 protein coding genes common in both were subjected to IPA. The results are listed in Table 24.

We particularly focused on the cancer or immuno-related biological functions (Figure 16). Figure 16A illustrates the top diseases and biological functions significantly identified by the common set of DEGs. These are categorized as Cancer, Organismal Injury and Abnormalities, and Cell Cycle. Figure 16B highlights the top biological functions from the unique set of DEGs in TNBC. These functions include Tissue Morphology, Cell Signaling, Immune Cell Trafficking and Inflammatory Response, Humoral Immune Response, Cell-mediated Immune Response, Cellular Movement and Development, Cellular Growth and Proliferation, Organismal Development and Morphology, and Cell Death and Survival. Figure16C highlights the top biological functions unique in ER+HER2-BC that are associated with Cellular Movement and Development, Cellular Growth and Proliferation, Tissue and Organismal Development, T-cell Signaling, Immune Cell Trafficking and Inflammatory Response. We noted that the functions of Cellular or Tissue Movement, and Cellar Growth and Proliferations identified from the unique set of DEGs in ER+HER2-BC were much more significant than the functions categorized in Inflammatory Response and Immune Response. This observation suggests these functions may play a dominant role during breast cancer cell development and growth for ER+HER2-BC. In contrast, the functions categorized as Cell Signaling, Immune Cell Trafficking and Inflammatory Response, Humoral Immune Response etc. identified from the unique set of DEGs in TNBC were much more significant than the functions categorized as Cellar Growth and Proliferations, and Cell Death and Survival. This observation suggests Inflammatory Response or Cell-mediated

Immune Response may play a dominant role for helping metastasis, which may be as a potential mechanism to explain why the TNBC patients has a poor survival rate.

We further examined the cancer-related genes in each category**.** Among the 896 genes in the common set, we observed 410 genes in Cancer, and 503 genes in Organismal Injury and Abnormalities. We found 389 genes in these two categories are associated with cancer; 128 genes are associated with BC; and 31 genes are associated with estrogen negative BC, 30 genes are associated with HER2- hormone receptor negative BC and 31 genes are associated with HER2-BC.

Among the 797 genes uniquely identified in TNBC, 294 and 381 genes are categorized in Cancer, and Organismal Injury and Abnormalities, respectively. We found 282 of these genes are associated with cancer and 135 genes are associated with breast cancer or colorectal cancer.

Among the 1403 genes uniquely identified by ER+HER2-BC, we found 498 genes in Cancer, 609 genes in Organismal Injury and Abnormalities, and 278 genes in Cellular growth and proliferation. We found 460 genes are associated with cancer; 223 genes are associated with BC or colorectal cancer, and 172 genes are associated with BC or ovarian cancer. More interestingly, we identified additional 26 genes besides 30 genes in common set that are also associated with HER2- hormone receptor negative BC. These genes could be potential biomarkers for the diagnosis of ER+HER2-BC subtype.

Table 24: The number of DEGs are associated with BC or other cancer with the aids of IPA.

	DEGs	Cancer	BC	BC or the other	ER-BC	HER2-hormone negative BC	HER2-BC
common in TNBC and ER+HER2-BC	896	389	128	198 (BC or CC) 159 (BC or OC)	31 (↑9, ↓21)	31(↑9, ↓22)	30(↑9, ↓21)
Unique in TNBC	797	282	-	135 (BC or OC)	-	-	-
Unique in ER+HER2-BC	1403	460	-	223 (BC or CC) 172 (BC or OC)	-	26(↑6, ↓20)	

Note: Breast Cancer (BC), Colorectal Cancer (CC), Ovarian Cancer (OC), ER negative (ER-), HER2 negative (HER2-).

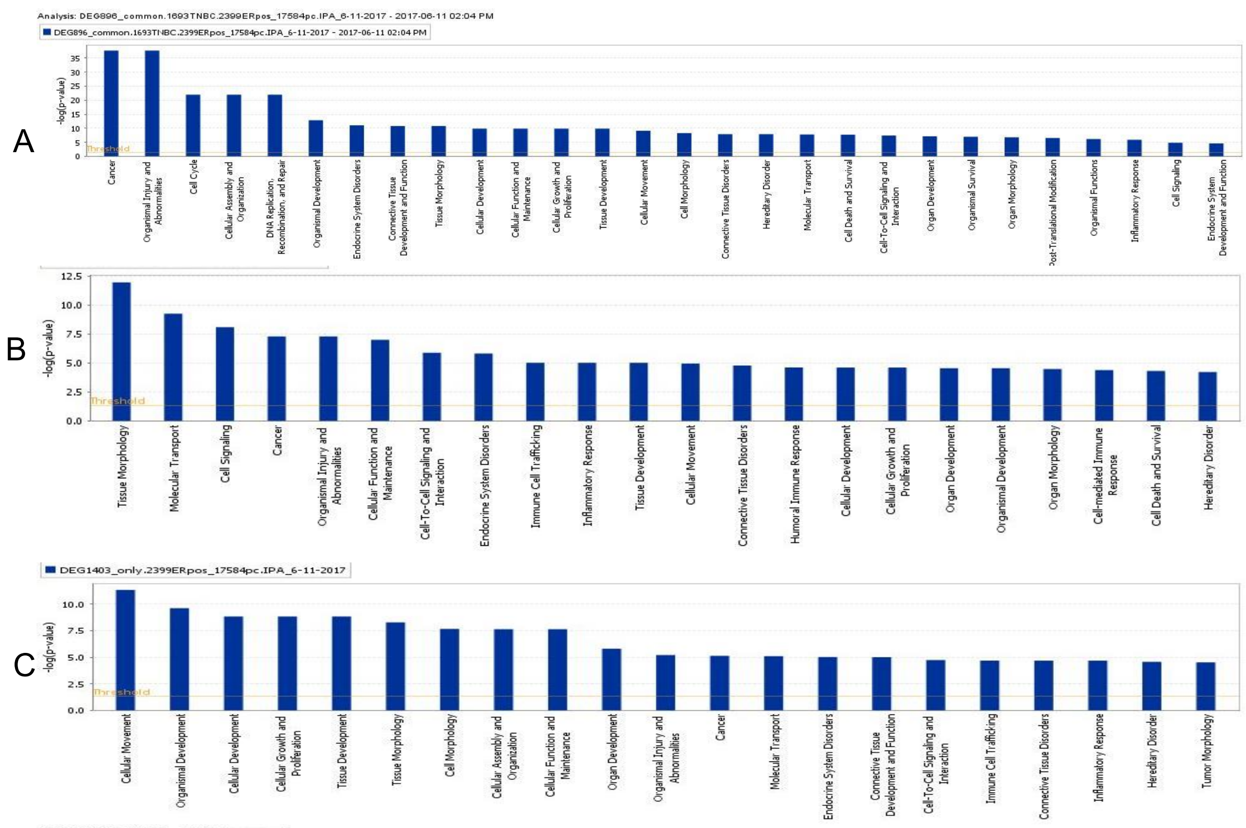

Figure 16: Canonical pathways identified from the common and unique DEGs in TNBC and ER+HER2-BC. A. 896 DEGs are common in TNBC and ER+HER2-BC. B. 797 DEGs unique in TNBC. C. 1403 DEGs unique in ER+HER2-BC.

7. Unique canonical pathways for TNBC and ER+HER2-BC

We further examined the canonical pathways and biomarkers using the 797 DEGs unique in TNBC and the1403 DEGs unique in ER+HER2-BC. For TNBC, we identified four cancer-related pathways including Estrogen Mediated S-phase Entry, cAMP-mediated Signaling, Calcium Signaling and LXR/RXR Activation (Figure 17A). These pathways play important roles in the regulation of cell cycle, promoting cell growth and proliferation or survival and apoptosis, and cell signaling. For example, cAMP-mediated intracellular signaling activates *ERK* via *EPAC1*, while *Src* and *Stat3* are activated by *Gai* and *Gao*. More interestingly, we identified several immuno-related signaling pathways including Altered T Cell and B cell Signaling, T helper Cell Differentiation, Complement System, B Cell Development, Agranulocyte Adhesion and Diapedesis, Intrinsic Prothrombin Activation Pathway and Clathrin-mediated Endocytosis Signaling (Figure 17B). These pathways play a crucial role in regulating numerous biological processes including cell growth, differentiation, survival, proliferation and metabolism via cellular immune response.

For ER+HER2-BC, we have identified many significant cancer-related pathways that were not found in TNBC including Notch Signaling, *FAK* Signaling, *ILK* Signaling, HER-2 Signaling in Breast Cancer, *PAK* Signaling, Paxillin Signaling, and Wnt/Ca^{+} Signaling, *ERK*/*MAPK* Signaling, and PCP Signaling (Figure 17B). These signaling pathways are associated with cellular growth, proliferation and organismal development. For example, Wnt/Ca^{+} Signaling is involved in various aspects of cell development like cell differentiation, growth and proliferation. Again, we noted that a fewer number of the pathways associated with cellular immune response were identified including

Granulocyte Adhesion and Diapedesis and Agranulocyte Adhesion and Diapedesis. Theses pathways may play a role in helping ER+HER2-BC to grow and penetration via an inflammatory response as the report in the recent studies.

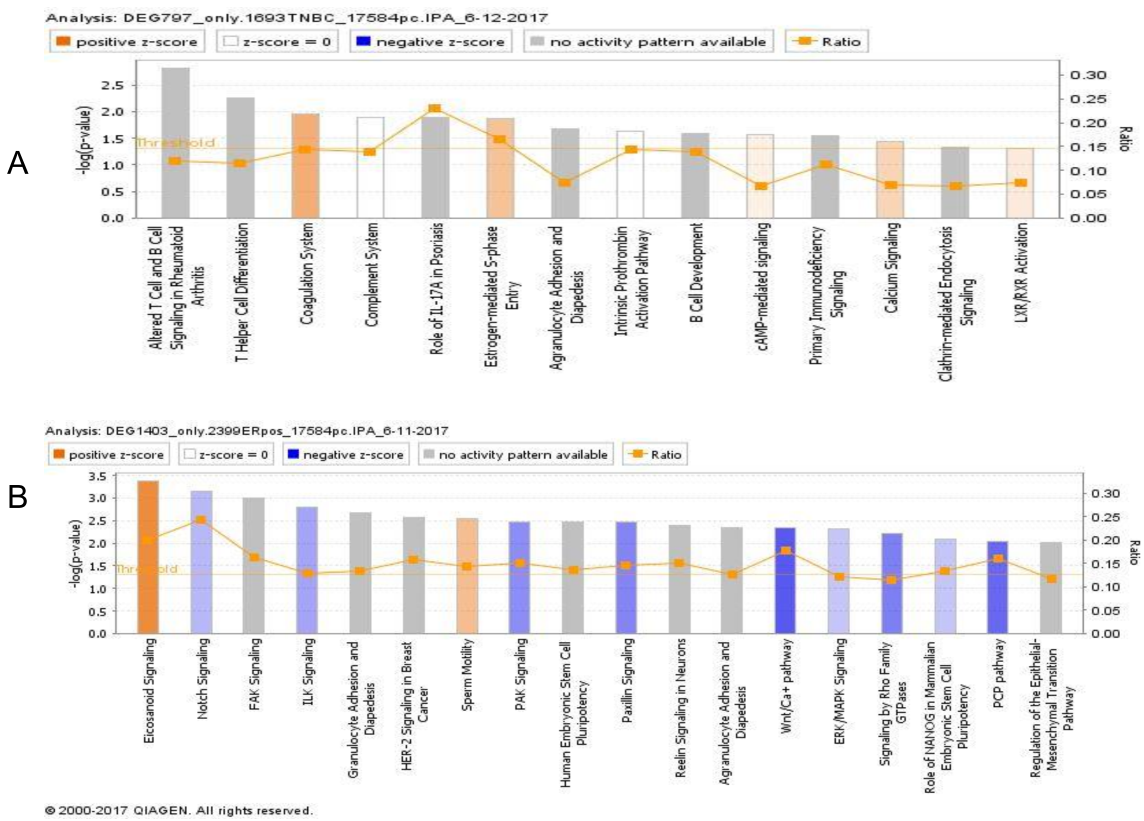

Figure 17: Canonical pathways identified from the unique DEGs in TNBC and ER+HER2-BC.

4. Summary and conclusion

It is known that a type I error is usually considered to be a more important and /or serious error which one would like to avoid [109]. An approach in a hypothesis for testing DEGs in the analysis of RNA-seq is to control type I error rate of $\boldsymbol{\alpha}$ using a nominal FDR at an acceptable level. Although the current existing methods for the DEGs analysis have taken into consideration of using a 0.05 FDR for the correction of the multiple genes, the studies in a comparison of all the methods including the commonly used DESeq and edegR methods report that they fail to maintain the actual FDR below the 0.05 nominal FDR resulting an inflated type I error rate [69, 78]. Most recently, a study in a comparison of normalization methods (DESeq in DESeq2, TMM in edgeR, FQ, Med-pgQ2 and UQ-pgQ2) using MAQC2 and MAQC3 observed that overall Med-pgQ2 and UQ-pgQ2 perform best by achieving a smaller actual FDR and higher specificity rate while maintain a sensitivity rate greater than 90% [37]. But this study also report that the DESeq normalization in DESeq2 perform best in terms of achieving an actual FDR, specificity and sensitivity at a quantile cutoff of the mean read counts less than 75% percentile. These studies suggest that Med-pgQ2 or UQ-pgQ2 are relatively conservative for high read counts of genes and DESeq in DESeq2 are relatively conservative for the gene expression below 75 percentile.

In this study, we use a combinational and approach to identify true DEGs using the public available BC datasets. Since the true DEGs are unknown and a sensitivity rate is unable to calculated, the optimal methods for this study will be mainly based on the FP genes identified using our approach. Although this is a limitation in this study, the previous study has shown the sensitivity rates for the DEGs analysis of MAQC2 and MAQC3 from DESeq2, edgeR and Med-pgQ2 or UQ-pgQ2methods were more than 90% given a 0.05 nominal FDR [37]. Referred to the previous studies, we first choose DESeq2, edgeR and UQ-pgQ2 with the aid of DESeq2 to perform the following DEGs analysis. We demonstrate that UQ-pgQ2 and *DESeq2* perform better than *edgeR* in terms of controlling FP genes identified via four with-group comparisons (two BC and two control groups).We observe that UQ-pgQ2 performs slightly better than DESeq2 with a FPR below 0.01. Furthermore, based on the FPR obtained from within group analysis, an optimal |logFC| cutoff is determined, which is used for the gene expression profiling.

Gene-expression profiling analysis has been used to dissect the heterogeneity of BC into six subtypes: Lumina A (ER+, low grade), Lumina B (ER+; high grade), HER2 positive (HER2-amplification), basa-like (ER−; HR−; HER2 −), normal-like and most recent “claudin low” subtypes [110-113]. Our results analyzing the gene expression profiles of two subtypes (TNBC and ER+HER2-BC) show their gene signatures are significantly different. We have identified 1693 true DEGs (protein coding) with a |logFC| ≥ 2 in TNBC versus control and 872 unique genes that are not identified in ER+HER2-BC. We have also identified 2299 DEGs (protein coding) with a |logFC| ≥ 1.5 in ER+HER2-BC versus control and 1042 genes unique in ER+HER2-BC. With the aid of IPA, these DEGs are categorized in Cancer, and Organismal Injury and Abnormalities among the top diseases and biological functions. For each unique set of genes, we further examine the genes associated with cancer or breast cancer. For the pathway analysis, we have also identified unique pathways of each set that are associated with cancer cell growth, proliferation and development or are involved in cellular immune response. These cancer-related DEGs may serve as potential biomarkers for the diagnosis of BC, BC subtype or potential targets for the immunotherapy treatment.

REFERENCES

1. Ghosh A, Islam T: **Genome-wide analysis and expression profiling of glyoxalase gene families in soybean (Glycine max) indicate their development and abiotic stress specific response**. *BMC plant biology* 2016, **16**(1):87.
2. Jiang Y, Malouf GG, Zhang J, Zheng X, Chen Y, Thompson EJ, Weinstein JN, Yuan Y, Spano JP, Broaddus R *et al*: **Long non-coding RNA profiling links subgroup classification of endometrioid endometrial carcinomas with trithorax and polycomb complex aberrations**. *Oncotarget* 2015, **6**(37):39865-39876.
3. Peffers MJ, Liu X, Clegg PD: **Transcriptomic profiling of cartilage ageing**. *Genomics data* 2014, **2**:27-28.
4. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al*: **De novo assembly and analysis of RNA-seq data**. *Nature methods* 2010, **7**(11):909-912.
5. Yue YJ, Liu JB, Yang M, Han JL, Guo TT, Guo J, Feng RL, Yang BH: **De novo assembly and characterization of skin transcriptome using RNAseq in sheep (Ovis aries)**. *Genetics and molecular research : GMR* 2015, **14**(1):1371-1384.
6. Schliebner I, Becher R, Hempel M, Deising HB, Horbach R: **New gene models and alternative splicing in the maize pathogen Colletotrichum graminicola revealed by RNA-Seq analysis**. *BMC genomics* 2014, **15**(1):842.
7. Wen L, He K, Yang H, Ni Y, Zhang X, Guo R, Pan Q: **Complete nucleotide sequence of a novel porcine circovirus-like agent and its infectivity in vitro**. *Science in China Series C, Life sciences / Chinese Academy of Sciences* 2008, **51**(5):453-458.
8. Craven KE, Gore J, Wilson JL, Korc M: **Angiogenic gene signature in human pancreatic cancer correlates with TGF-beta and inflammatory transcriptomes**. *Oncotarget* 2016, **7**(1):323-341.
9. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
10. Voelkerding KV, Lyon E: **Digital fetal aneuploidy diagnosis by next-generation sequencing**. *Clinical chemistry* 2010, **56**(3):336-338.
11. Hansen KD, Wu Z, Irizarry RA, Leek JT: **Sequencing technology does not eliminate biological variability**. *Nature biotechnology* 2011, **29**(7):572-573.
12. Auer PL, Doerge RW: **Statistical design and analysis of RNA sequencing data**. *Genetics* 2010, **185**(2):405-416.
13. Fang Z, Cui X: **Design and validation issues in RNA-seq experiments**. *Briefings in bioinformatics* 2011, **12**(3):280-287.
14. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome research* 2008, **18**(9):1509-1517.
15. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
16. Li CI, Su PF, Guo Y, Shyr Y: **Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution**. *International journal of computational biology and drug design* 2013, **6**(4):358-375.

17. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome biology* 2010, **11**(10):R106.
18. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
19. Li CI, Su PF, Shyr Y: **Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data**. *BMC bioinformatics* 2013, **14**:357.
20. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing**. *JRStatist soc B* 1995, **57**:289-300.
21. Storey JD: **A direct approach to false discovery rates**. *J R Stat Soc Ser B* 2002, **64**:479-498.
22. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data**. *Biostatistics* 2008, **9**(2):321-332.
23. Aban IB, Cutter GR, Mavinga N: **Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data**. *Computational statistics & data analysis* 2008, **53**(3):820-833.
24. Ng HK, Tang ML: **Testing the equality of two Poisson means using the rate ratio**. *Statistics in medicine* 2005, **24**(6):955-965.
25. Gu K, Ng HK, Tang ML, Schucany WR: **Testing the ratio of two poisson rates**. *Biometrical journal Biometrische Zeitschrift* 2008, **50**(2):283-298.
26. Thode HC: **Power and sample size requirements for tests of differences between two LPoisson rates.** *The Statistician* 1997, **46**:227-230.
27. Krishnamoorhy K, Thomson J: **A more powerful test for comparing two poisson means.** *J Stat Plan Infer* 2004, **119**:23-35.
28. Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK, Nesmith AS, Oliver PG, Grizzle WE, Forero A *et al*: **Recurrent read-through fusion transcripts in breast cancer**. *Breast cancer research and treatment* 2014, **146**(2):287-297.
29. Wright HL, Cox T, Moots RJ, Edwards SW: **Neutrophil biomarkers predict response to therapy with tumor necrosis factor inhibitors in rheumatoid arthritis**. *Journal of leukocyte biology* 2016.
30. Xu L, Ziegelbauer J, Wang R, Wu WW, Shen RF, Juhl H, Zhang Y, Rosenberg A: **Distinct Profiles for Mitochondrial t-RNAs and Small Nucleolar RNAs in Locally Invasive and Metastatic Colorectal Cancer**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2016, **22**(3):773-784.
31. Guo W, Wang Q, Zhan Y, Chen X, Yu Q, Zhang J, Wang Y, Xu XJ, Zhu L: **Transcriptome sequencing uncovers a three-long noncoding RNA signature in predicting breast cancer survival**. *Scientific reports* 2016, **6**:27931.
32. Smith BA, Sokolov A, Uzunangelov V, Baertsch R, Newton Y, Graim K, Mathis C, Cheng D, Stuart JM, Witte ON: **A basal stem cell signature identifies aggressive prostate cancer phenotypes**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(47):E6544-6552.
33. Yao F, Zhang C, Du W, Liu C, Xu Y: **Identification of Gene-Expression Signatures and Protein Markers for Breast Cancer Grading and Staging**. *PloS one* 2015, **10**(9):e0138213.
34. Rajan P, Stockley J, Sudbery IM, Fleming JT, Hedley A, Kalna G, Sims D, Ponting CP, Heger A, Robson CN *et al*: **Identification of a candidate prognostic gene signature by transcriptome analysis of matched pre- and post-treatment prostatic biopsies from patients with advanced prostate cancer**. *BMC cancer* 2014, **14**:977.

35. Wright HL, Thomas HB, Moots RJ, Edwards SW: **Interferon gene expression signature in rheumatoid arthritis neutrophils correlates with a good response to TNFi therapy**. *Rheumatology* 2015, **54**(1):188-193.
36. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome biology* 2010, **11**(3):R25.
37. Li X, Cooper GF, Shyr Y, Wu D, Rouchka EC, Gill RS, O'Toole TE, Brock GN, Rai SN: **Inference and Sample Size Calculations Based on Statistical Tests in a Negativ ebnomial Distribution for Differential Gene Expression in RNA-seq Data**. *J Biom Biostat* 2017, **8**(1).
38. Keene ON, Jones MR, Lane PW, Anderson J: **Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the TRISTAN study**. *Pharmaceutical statistics* 2007, **6**(2):89-97.
39. Zhu H, Lakkis H: **Sample size calculation for comparing two negative binomial rates**. *Statistics in medicine* 2014, **33**(3):376-387.
40. Mi G, Di Y, Schafer DW: **Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data**. *PloS one* 2015, **10**(3):e0119254.
41. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nature genetics* 2008, **40**(12):1413-1415.
42. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels**. *Bioinformatics* 2012, **28**(8):1086-1092.
43. Wu P, Zhang H, Lin W, Hao Y, Ren L, Zhang C, Li N, Wei H, Jiang Y, He F: **Discovery of novel genes and gene isoforms by integrating transcriptomic and proteomic profiling from mouse liver**. *Journal of proteome research* 2014, **13**(5):2409-2419.
44. Canovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF: **SNP discovery in the bovine milk transcriptome using RNA-Seq technology**. *Mammalian genome : official journal of the International Mammalian Genome Society* 2010, **21**(11-12):592-598.
45. Djari A, Esquerre D, Weiss B, Martins F, Meersseman C, Boussaha M, Klopp C, Rocha D: **Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing**. *BMC genomics* 2013, **14**:307.
46. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data**. *American journal of human genetics* 2013, **93**(4):641-651.
47. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.
48. Marguerat S, Bahler J: **RNA-seq: from technology to biology**. *Cellular and molecular life sciences : CMLS* 2010, **67**(4):569-579.
49. Voelkerding KV, Dames SA, Durtschi JD: **Next-generation sequencing: from basic research to diagnostics**. *Clinical chemistry* 2009, **55**(4):641-658.
50. Chrystoja CC, Diamandis EP: **Whole genome sequencing as a diagnostic test: challenges and opportunities**. *Clinical chemistry* 2014, **60**(5):724-733.
51. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(14):8418-8423.
52. Zeng W, Mortazavi A: **Technical considerations for functional sequencing assays**. *Nature immunology* 2012, **13**(9):802-807.

53. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.
54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
55. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.
56. Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R: **Evaluation of normalization methods for microarray data**. *BMC bioinformatics* 2003, **4**:33.
57. Quackenbush J: **Microarray data normalization and transformation**. *Nature genetics* 2002, **32 Suppl**:496-501.
58. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC bioinformatics* 2010, **11**:94.
59. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
60. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data**. *Genome biology* 2013, **14**(9):R95.
61. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods* 2008, **5**(7):621-628.
62. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology**. *Biology direct* 2009, **4**:14.
63. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al*: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Briefings in bioinformatics* 2013, **14**(6):671-683.
64. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.
65. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature biotechnology* 2010, **28**(5):511-515.
66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature protocols* 2012, **7**(3):562-578.
67. Risso D, Ngai J, Speed TP, Dudoit S: **Normalization of RNA-seq data using factor analysis of control genes or samples**. *Nature biotechnology* 2014, **32**(9):896-902.
68. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-Seq data**. *BMC bioinformatics* 2011, **12**:480.
69. Kvam VM, Liu P, Si Y: **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data**. *American journal of botany* 2012, **99**(2):248-256.
70. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Briefings in bioinformatics* 2013.

71. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC bioinformatics* 2013, **14**:91.
72. Brunner AL, Li J, Guo X, Sweeney RT, Varma S, Zhu SX, Li R, Tibshirani R, West RB: **A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions**. *Genome biology* 2014, **15**(5):R71.
73. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
74. Yu D, Huber W, Vitek O: **Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size**. *Bioinformatics* 2013, **29**(10):1275-1282.
75. Cameron AC, Trivedi PK: **Regression analysis of count data**. *Cambridge University Press* 1998:566.
76. Consortium M, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES *et al*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nature biotechnology* 2006, **24**(9):1151-1161.
77. Wan L, Sun F: **CEDER: accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq**. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 2012, **9**(5):1281-1292.
78. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR *et al*: **A comparative study of techniques for differential expression analysis on RNA-Seq data**. *PloS one* 2014, **9**(8):e103207.
79. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nature biotechnology* 2013, **31**(1):46-53.
80. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome biology* 2014, **15**(12):550.
81. Li X, Brock GN, Rouchka EC, Cooper NGF, Wu D, O'Toole TE, Gill RS, Eteleeb AM, O'Brien L, Rai SN: **A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data**. *PloS one* 2017, **12**(5):e0176185.
82. Fawcett T: **An introduction to ROC analysis**. *Pattern Recognition Letters* 2006, **27**:861-874.
83. Hanley JA, McNeil BJ: **The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve**. *Radiology* 1982, **143**:29-36.
84. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Briefings in bioinformatics* 2015, **16**(1):59-70.
85. Ness SA: **Basic microarray analysis: strategies for successful experiments**. *Methods in molecular biology* 2006, **316**:13-33.
86. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.
87. Becker C, Hammerle-Fickinger A, Riedmaier I, Pfaffl MW: **mRNA and microRNA quality control for RT-qPCR analysis**. *Methods* 2010, **50**(4):237-243.
88. Walleshauser III JG, Kessler T, Morse D, Tannous BA, Chiu NHL: **A Simple Approach for Evaluating Total MicroRNA Extraction from Mouse brain Tissues**. *Journal of Analytical Sciences, Methods and Instrumentation* 2012, **2**:5-12.

89. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome biology* 2014, **15**(2):R29.
90. Schroeder RL, Stevens CL, Sridhar J: **Small molecule tyrosine kinase inhibitors of ErbB2/HER2/Neu in the treatment of aggressive breast cancer**. *Molecules* 2014, **19**(9):15196-15212.
91. Zhang L, Huang Y, Zhuo W, Zhu Y, Zhu B, Chen Z: **Identification and characterization of biomarkers and their functions for Lapatinib-resistant breast cancer**. *Medical oncology* 2017, **34**(5):89.
92. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, Forman D, Bray F: **Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012**. *European journal of cancer* 2013, **49**(6):1374-1403.
93. El Saghir NS, Khalil MK, Eid T, El Kinge AR, Charafeddine M, Geara F, Seoud M, Shamseddine AI: **Trends in epidemiology and management of breast cancer in developing Arab countries: a literature and registry analysis**. *International journal of surgery* 2007, **5**(4):225-233.
94. Makoukji J, Makhoul NJ, Khalil M, El-Sitt S, Aldin ES, Jabbour M, Boulos F, Gadaleta E, Sangaralingam A, Chelala C *et al*: **Gene expression profiling of breast cancer in Lebanese women**. *Scientific reports* 2016, **6**:36639.
95. Lee BL, Liedke PE, Barrios CH, Simon SD, Finkelstein DM, Goss PE: **Breast cancer in Brazil: present status and future goals**. *The Lancet Oncology* 2012, **13**(3):e95-e102.
96. Kwei KA, Kung Y, Salari K, Holcomb IN, Pollack JR: **Genomic instability in breast cancer: pathogenesis and clinical implications**. *Molecular oncology* 2010, **4**(3):255-266.
97. Matta J, Morales L, Ortiz C, Adams D, Vargas W, Casbas P, Dutil J, Echenique M, Suarez E: **Estrogen Receptor Expression Is Associated with DNA Repair Capacity in Breast Cancer**. *PloS one* 2016, **11**(3):e0152422.
98. Martin HL, Smith L, Tomlinson DC: **Multidrug-resistant breast cancer: current perspectives**. *Breast cancer* 2014, **6**:1-13.
99. Raha P, Thomas S, Munster PN: **Epigenetic modulation: a novel therapeutic target for overcoming hormonal therapy resistance**. *Epigenomics* 2011, **3**(4):451-470.
100. Sutherland RL: **Endocrine resistance in breast cancer: new roles for ErbB3 and ErbB4**. *Breast cancer research : BCR* 2011, **13**(3):106.
101. Zhang MH, Man HT, Zhao XD, Dong N, Ma SL: **Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review)**. *Biomedical reports* 2014, **2**(1):41-52.
102. Garcia-Becerra R, Santos N, Diaz L, Camacho J: **Mechanisms of resistance to endocrine therapy in breast cancer: focus on signaling pathways, miRNAs and genetically based resistance**. *International journal of molecular sciences* 2012, **14**(1):108-145.
103. Malzahn K, Mitze M, Thoenes M, Moll R: **Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas**. *Virchows Archiv : an international journal of pathology* 1998, **433**(2):119-129.
104. Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M *et al*: **Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008, **26**(8):1275-1281.
105. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM: **The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2007, **13**(8):2329-2334.

106. Podo F, Buydens LM, Degani H, Hilhorst R, Klipp E, Gribbestad IS, Van Huffel S, van Laarhoven HW, Luts J, Monleon D *et al*: **Triple-negative breast cancer: present challenges and new perspectives**. *Molecular oncology* 2010, **4**(3):209-229.
107. von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J *et al*: **Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2012, **30**(15):1796-1804.
108. Al-Ejeh F, Simpson PT, Saunus JM, Klein K, Kalimutho M, Shi W, Miranda M, Kutasovic J, Raghavendra A, Madore J *et al*: **Meta-analysis of the global gene expression profile of triple-negative breast cancer identifies genes for the prognostication and treatment of aggressive breast cancer**. *Oncogenesis* 2014, **3**:e124.
109. Chow SC, Shao J, Wang H: **Sample Size calcualtions in Clinical Research**. 2003.
110. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.
111. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10869-10874.
112. Hennessy BT, Gonzalez-Angulo AM, Stemke-Hale K, Gilcrease MZ, Krishnamurthy S, Lee JS, Fridlyand J, Sahin A, Agarwal R, Joy C *et al*: **Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics**. *Cancer research* 2009, **69**(10):4116-4124.
113. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L *et al*: **The molecular portraits of breast tumors are conserved across microarray platforms**. *BMC genomics* 2006, **7**:96.

APPENDIX

Part of this dissertation has been previously published. This includes:

1) The content, Figures and Tables for the subsection entitled “SAMPLESIZE CALCULATION METHODS BASED ON STATISTICAL TESTS IN A NEGATIVE BILNOMIAL DISTRIBUTION FOR RNA-SEQ DATA” was published in the Journal of Biometrics and Biostatistics, 2017, 8(1) [37].

2) The content, and Figures and Tables for the subsection entitled “NORMALIZATION METHODS FOR RNA-SEQ DATA” was published in PLOS ONE, 2017, 12(5):e0176185 [81].

CURRICULUM VITA

1. CONTACT INFORMATION

Xiaohong Li

Department of Bioinformatics and Biostatistics

Bioinformatics Core

University of Louisville

Louisville, KY, 40202

Phone 502-852-1809

Xiaohong.Li@lousiville.edu

2. CITIZENSHIP

Neutralized Canadian and American

3. ACADEMIC DEGREE

2011-2017 PhD Candidate (**GPA 3.95**)

Department of Bioinformatics and Biostatistics

University of Louisville

2008-2010 Master of Science in Bioinformatics and Biostatistics (**GPA 3.95**)

Department of Bioinformatics and Biostatistics

University of Louisville

1998-2002 B.S. in Computer and Information Sciences (**Cum Laude**)

School of Business

Cleveland State University

1998-1999 Upper level undergraduate student in Computer Science

The Institute of Technology

University of Minnesota

4. HONORS

1998-2002 Golden Key International Honor Society for excellence in undergraduate programs

2000-2002 Beta Gamma Sigma, International Honor Society for outstanding academic achievements of business students

2008-present Golden Key Honor Society for excellence in graduate programs

5. Working experience

June 2016-present Biostatistician II in Bioinformatics core

Routinely conducting genomic data analysis of data such as RNA-seq, Chip-seq and Bacteria 16S rRNA gene sequencing data generated in Genomic core in University of Louisville, and writing a report to the researchers. Initiating and exploring the new methods for the new datasets. Providing service to assist the researchers for their research studies and grant applications. Participate in Bioinformatics club and KBRIN summit.

Oct. 2013- 2016 Biostatistician I in Bioinformatics core

University of Louisville

Routinely conducting genomic data analysis such as RNA-seq, Chip-seq and 16S data generated in Genomic core and writing a report to the researchers.

Sept. 2003-2014 Senior research technologist in DNA sequencing core

University of Louisville

In DNA sequencing core, generating and analyze Sanger DNA sequencing and write a report to the researchers. Generating and analyzing Agilent microarray with variety of species and reporting the results to the researchers. Generating DNA-seq data using ion torrent platform.

Jan. 2000-2002 Research technologist in Dept. of Molecular Cardiology

	Cleveland Clinic Foundation
	Assist senior scientist to perform an experiment. Routine work includes DNA cloning, analyzing DNA-seq results, RT-PCR, cell culture and western blotting.
Feb. 1998-1999	Junior Scientist in Dept. of Pharmacology
	University of Minnesota
	Routinely perform DNA cloning, RT-PCR, cell culture, Southern blot and DNA sequencing.
Aug. 1995-1997	Research assistant in Dept. of Veterinary Microbiology
	University of Saskatchewan
	Assist graduate students to conduct an experiment and teach the students the basic laboratory skills (cell culture, RT-PCR, Northern Blot and Western Blot etc.) that are needed in their study.
May1988-1990	Instructor in Dept. of nursing
	Yangtze University

6. PUBLICATION

Book Chapter

Xiaohong Li, Carolyn M. Klinge, Susmita Datta. Novel and Alternative Bioinformatics Approaches to Understand microRNA-mRNA Interactome in Cancer Research**. In Systems Biology in Cancer Research and Drug Discovery**. 2012, Part 3, 267-288, DOI: 10.1007/978-94-007-4819-4_11

Journal Articles

1. Gordon J.R, Cockcroft D.W, **Li XH.** Cytokine expression on cross-linking of human tissue Fc Epsilon RI. Int. Arch Allergy Immunol 113 (1997): 269-271.
2. Kendall JC, **Li XH**, Galli SJ, and Gordon JR. Promotion of mouse fibroblast proliferation by IgE-dependent activation of mouse mast cells: role for mast cell tumor necrosis factor-a and transforming growth factor-b1. J. Allergy Clin. Immunol. 99(1997): 113-123.
3. Wei Fan, **Xiaohong Li** and Nigel G. F. Cooper: CaMKIIaB Mediates a Survival Response in Retinal Ganglion Cell Subjected to a Glutamate Stimulus. Investigative Ophthalmology and Visual Science. 48(2007):3854-3863
4. Mark E. Bardgett*[1], Brian Hoffman[1], Molly S. Griffith[1], **Xiaohong Li**[2], & Nigel G. F. Cooper[2]. Global gene expression profiling of rat prefrontal cortex after

chronic oral risperidone treatment. Abstract, 41st Winter Conference on Brain Research-Snowbird, Utah, 2008

5. W. Fan, **X. Li,** W. Wang, J.S.Mo, H. Kaplan and N.G.F. Cooper. Early Involvement of Immune/Inflammatory Response Genes in Retinal Degradation in DBA/2J mice (http://www.la-press.com). Ophthalmology and Eye Diseases. 1(2009):1:19
6. Julie B Schuck, Chia-Hui Lin, William T Penberthy, **Xiaohong Li**, Nigel GF Cooper and Michael E Smith. Microarray analysis and quantitative real-time PCR validation of gene expression during auditory hair cell regeneration in zebrafish(*Danio rerio*). BMC Bioinformatics 10(2009): A12
7. Eng M, Brock G, **Li X,** Chen Y, Ravindra KV, Buell JF and Marvin MR. Perioperative anticoagulation and antiplatelet therapy in renal transplant: is there an increase in bleeding complication. Clin Transplant 2010.
8. Timothy E. O'Toole, Katarzyna Bialkowska, **Xiaohong Li,** and Joan E.B. Fox: Tiam1 is Recruited to β1-Integrain Complexes by 14-3-3ζ Where it Mediates integrin-induced Rac1 Activation and Motility. J Cell Physiol 226:2965-78, 2011.
9. **Xiaohong Li**, Ryan Gill, Nigel G. F. Cooper, Jae Keun Yoo and Susmita Datta. Modeling microRNA-mRNA Interactions Using PLS Regression in Human Colon Cancer. BMC Med Genomics 4:44, 2011 (highly accessed).
10. Julie B Schuck, Huifang Sun, W Todd PenberthywT, Nigel GF Cooper, **Xiaohong Li,** Michael E Smith. Transcriptomic analysis of the zebrafish inner ear points to growth hormone mediated regeneration following acoustic trauma. BMC Neurosci. 12:88, 2011 (highly accessed).
11. Ye F, Yuan F, **Li X**, Cooper N, Tinney JP, Keller BB. Gene expression profiles in engineered cardiac tissues respond to mechanical loading and inhibition of tyrosine kinases. Physiol Rep. 1(5):e00078, 2013
12. O'Toole TE, Abplanalp W, **Li X**, Cooper N, Conklin DJ, Haberzetll P, and Bhatnagar A. Acrolein decreases endothelial cell migration and insulin sensitivity through induction of let-7a. Toxicol Sci. 140(2):271-82, 2014.
13. Andreeva K, Zhang M, Fan W, **Li X**, Chen Y, Rebolledo-Mendez JD, Cooper NG. Time-dependent Gene Profiling Indicates the Presence of Different Phases for Ischemia/Reperfusion Injury in Retina. Ophthalmol Eye Dis. 6:43-54, 2014
14. Jeoung M, Jang ER, Liu J, Wang c, Rouchka EC, **Li X**, Galperin E. Shoc2-tranduced ERK1/2 motility signals-Novel insights form functional genomics. Cell Signal. 28(5):448-59, 2016
15. Rouchka EC, Jeoung M, Jang ER, Liu J, Wang C, **Li X**, Galperin E. Data set for transcriptional response to depletion of the Shoc2 scaffolding protein. Data Brief. 7:770-8, 2016.
16. **Li X,** Cooper GF, Shyr Y, Wu D, Rouchka EC*, et al.* Inference and Sample Size Calculations Based on Statistical Tests in a Negativ ebnomial Distribution for Differential Gene Expression in RNA-seq Data. *J Biom Biostat* ; 8, 2017
17. **Li X**, Brock GN, Rouchka EC, Cooper NG, Wu D, O'Toole TE, Gill RS, Eteleeb AM, O'Brien L, Rai SN. A comparison of per Sample Global Scaling

and per Gene Normalization Methods for Differential Expression Analysis of RNA-seq Data. PLoS One. **12**(5):e0176185, 2017

18. Bamji SF, Rouchka E, Zhang Y, **Li X,** Kalbfleisch T, Corbitt C. Next generation sequencing analysis of soy glyceollins and 17-$\boldsymbol{\beta}$ estradiol: Effects on transcript abundance in the female mouse brain. Mol Cell Endocrinol. 1-7, 2017.

Submitted

Li X, Wu D, Cooper NG, Rai SN. Sample size calculations using a generalized linear model in RNA-seq data. Submitted into *Statistics in Medicine*.

In Preparation

Identification of optimal normalization and test statistical methods for differential gene expression analysis in human breast cancer drafted by Xiaohong Li