BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data

Chung-I Li[1,3], Pei-Fang Su[2,3] and Yu Shyr[3*]

Abstract

Background: Sample size calculation is an important issue in the experimental design of biomedical research. For RNA-seq experiments, the sample size calculation method based on the Poisson model has been proposed; however, when there are biological replicates, RNA-seq data could exhibit variation significantly greater than the mean (i.e. over-dispersion). The Poisson model cannot appropriately model the over-dispersion, and in such cases, the negative binomial model has been used as a natural extension of the Poisson model. Because the field currently lacks a sample size calculation method based on the negative binomial model for assessing differential expression analysis of RNA-seq data, we propose a method to calculate the sample size.

Results: We propose a sample size calculation method based on the exact test for assessing differential expression analysis of RNA-seq data.

Conclusions: The proposed sample size calculation method is straightforward and not computationally intensive. Simulation studies to evaluate the performance of the proposed sample size method are presented; the results indicate our method works well, with achievement of desired power.

Background

Next generation sequencing (NGS) technology has revolutionized genetic analysis; RNA-seq is a powerful NGS method that enables researchers to discover, profile, and quantify RNA transcripts across the entire transcriptome. In addition, unlike the microarray chip, which offers only quantification of gene expression level, RNA-seq provides expression level data as well as differentially spliced variants, gene fusion, and mutation profile data. Such advantages have gradually elevated RNA-seq as the technology of choice among researchers. Nevertheless, the advantages of RNA-seq are not without computational cost; as compared to microarray analysis, RNA-seq data analysis is much more complicated and difficult. In the past several years, the published literature has addressed the application of RNA-seq to multiple research questions, including abundance estimation [1-3], detection of alternative splicing [4-6], detection of novel transcripts [6,7], and the biology associated with gene expression profile differences between samples [8-10]. With this rapid growth of RNA-seq applications, discussion of experimental design issues has lagged behind, though more recent literature has begun to address some of the relevant principles (e.g., randomization, replication, and blocking) to guide decisions in the RNA-seq framework [11,12].

One of the principal questions in designing an RNA-seq experiment is: What is the optimal number of biological replicates to achieve desired statistical power? (Note: In this article, the term "sample size" is used to refer to the number of biological replicates or number of subjects.) Because RNA-seq data are counts, the Poisson distribution has been widely used to model the number of reads obtained for each gene to identify differential gene expression [8,13]. Further, [12] used a Poisson distribution to model RNA-seq data and derive a sample size calculation formula based on the Wald

*Correspondence: yu.shyr@vanderbilt.edu
[3]Center for Quantitative Sciences, Vanderbilt University, 571 Preston Building Nashville, TN, USA
Full list of author information is available at the end of the article

test for single-gene differential expression analysis. It is worth noting that a critical assumption of the Poisson model is that the mean and variance are equal. This assumption may not hold, however, as read counts could exhibit variation significantly greater than the mean [14]. That is, the data are over-dispersed relative to the Poisson model. In such cases, one natural alternative to Poisson is the negative binomial model. Based on the negative binomial model, [14,15] proposed a quantile-adjusted conditional maximum likelihood procedure to create a pseudocount which lead to the development of an exact test for assessing the differential expression analysis of RNA-seq data. Furthermore, [16] provided a Bioconductor package, edgeR, based on the exact test.

Sample size determination based on the exact test has not yet been studied, however. Therefore, the first goal of this paper is to propose a sample size calculation method based on the exact test.

In reality, thousands of genes are examined in an RNA-seq experiment; differential expression among those genes is tested simultaneously, requiring the correction of error rates for multiple comparisons. For the high-dimensional multiple testing problem, several such corrected measures have been proposed, such as family-wise error rate (FWER) and false discovery rate (FDR). In high-dimensional multiple testing circumstances, controlling FDR is preferable [17] because the Bonferroni correction for FWER is often too conservative [18]. Many methods have been proposed to control FDR in the analysis of high-dimensional data [17,19,20]. Those concepts have been extended to calculate sample size for microarray studies [21-25]. To our knowledge, however, the literature does not address determination of sample size while controlling FDR in RNA-seq data. Therefore, the second purpose of this paper is to propose a procedure to calculate sample size while controlling FDR for differential expression analysis of RNA-seq data.

In sum, in this article, we address the following two questions: (i) For a single-gene comparison, what is the minimum number of biological replicates needed to achieve a specified power for identifying differential gene expression between two groups? (ii) For multiple gene comparisons, what is the suitable sample size while controlling FDR? The article is organized as follows. In the Method section, a sample size calculation method is proposed for a single-gene comparison. We then extend the method to address the multiple comparison test issue. Performance comparisons via numerical studies are described in the Results section. Two real RNA-seq data sets are used to illustrate sample size calculation. Finally, discussion follows in the Conclusions section.

Method

Exact test

In an RNA-seq experiment, the total number of reads, also referred to as library size, mapped to the genome are different among the samples. In such cases, the counts in each group are not identically distributed, and it is difficult to develop an exact test for assessing the differential expression analysis of RNA-seq data. To handle this issue, [14,15] proposed a quantile-adjusted conditional maximum likelihood procedure to create pseudocounts which are approximately identically distributed and which lead to the development of an exact test. In the following, the proposed sample size calculation method is based the exact test for a single-gene comparison. Let Y_{ij} be the random variable corresponding to the pseudocount, with y_{ij} being the observed value of Y_{ij}, of the jth ($j = 1, 2, \ldots, n_i$) sample of the ith ($i = 0, 1$) group where n_0 and n_1 are the numbers of samples from the control and treatment group, respectively. Assume pseudocount Y_{ij} can be modeled as a negative binomial (NB) distribution, NB($d_{ij}\gamma_i, \phi$). Here, γ_i represents the normalized gene expression level of group i, d_{ij} represents a normalization factor for the total number of reads mapped in the jth sample of the ith group, and ϕ is the dispersion. We use the NB parameterization where the mean is $\mu_{ij} = d_{ij}\gamma_i$ and variance is $\mu_{ij}(1 + \mu_{ij}^2\phi)$. Because the question of interest is to identify the differential gene expression between two groups, the corresponding testing hypothesis is

$$H_0 : \gamma_1 = \gamma_0 \ \ vs. H_1 : \gamma_1 \neq \gamma_0. \quad (1)$$

Because the pseudocounts in each group have an approximately identical negative binomial distribution [14,15], the sum of pseudocounts of each group, $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, has a negative binomial distribution NB($n_i d_i^* \gamma_i, \phi/n_i$) where d_i^* is the geometric mean of normalization factors in group i. Under the null hypothesis (1), the sum of the total pseudocount, $Y_1 + Y_0$, follows a negative binomial distribution. In analogy with Fisher's exact test, [14,15] proposed an exact test for replacing the hypergeometric probabilities with negative binomial probabilities. Because [16] developed a Bioconductor software package edgeR which is an implementation of methodology developed by [14,15], the p-value can be easily calculated for conducting the exact test.

In the following simulation and application sections, we used edgeR version 3.0.6 for estimating model parameters and performing the exact test.

Sample size calculation for controlling type I error rate

In this section, we focus on sample size calculation based on the exact test for a single-gene comparison as described in the test statistics section. For simplicity, we assume the RNA-seq experiment uses a balanced design (i.e., $n_0 = n_1 = n$), which is a special but common case.

The following method could be easily extended to the unbalanced case (i.e. let $n_0 = n$ and $n_1 = kn$ where k is a predetermined ratio of the sample size of the control group to the treatment group). In order to perform sample size calculations, it is necessary to construct a power function for the testing described above. The power of a test is the probability that the null hypothesis is rejected when the alternative hypothesis is true. Since the distribution of the exact test statistic under the alternative hypothesis is unknown, however, it is difficult to derive a closed-form expression of the power function. Instead of deriving the distribution of test statistic under the alternative hypothesis, [26] proposed a method to calculate the power for the exact test based on a given p-value. Here, we borrow their concept to calculate power. For a given p-value, $p(y_1, y_0)$ where y_0 and y_1 are the observed pseudo-sums, described in the previous section, the power can be expressed as

$$\xi(n,\rho,\mu_0,\phi,w,\alpha) = \sum_{y_0=0}^{\infty}\sum_{y_1=0}^{\infty} f\left(nw\rho\mu_0, \frac{\phi}{n}\right) f\left(n\mu_0, \frac{\phi}{n}\right) I(p(y_1,y_0) < \alpha),$$

where $w = d_1^*/d_0^*$ is the ratio of the geometric means of normalization factors between two groups, $\rho = \gamma_1/\gamma_0$ is the fold change, $\mu_0 = d_0^*\gamma_0$ is the average number of reads in the control group, $f(\mu,\phi)$ is the probability mass function of the negative binomial distribution with mean μ as well as dispersion ϕ, α is the the level of significance, and $I(.)$ denotes the indicator function. For a given desired power $1-\beta$, the power of the test can be represented as the function of sample size in the form

$$1 - \beta = \xi(n,\rho,\mu_0,\phi,w,\alpha). \quad (2)$$

Thus, the required sample size n to attain the given power $1-\beta$ at level of significance α can then be calculated by solving (2) through a numerical approach, such as a gradient-search or bisection procedure.

Sample size calculation for controlling false discovery rate

In reality, thousands of genes are examined in an RNA-seq experiment, and those genes are tested simultaneously for significance of differential expression. In such cases, the sample size calculation for a single-gene comparison discussed above cannot be applied directly. Jung, 2005 [23] incorporated FDR controlling based on a two-sample t-test under the Gaussian distribution assumption. In this section, we borrowed their concept to incorporate FDR controlling based on the test statistics described in the test statistics section.

For the multiple testing problem, [19] suggested the use of false discovery rate (FDR) which is defined as the expected proportion of false discoveries among rejected null hypotheses. Storey, 2002 [17] further proposed an improvement to FDR to achieve higher power, in the form

$$\mathrm{FDR} = E\left(\frac{R_0}{R} \;\middle|\; R > 0\right),$$

where R_0 is the number of false discoveries and R is the number of results declared significant (i.e., rejections of the null hypothesis).

To calculate the sample size for microarray data analysis, [23] proposed an FDR-controlled method which is based on the expression of FDR under independence (or weak dependence) among test statistics, as

$$\mathrm{FDR} = \frac{m_0\alpha}{m_0\alpha + E(R_1)},$$

[17,27], where m_0 is the number of true null hypotheses and $E(R_1)$ is the expected number of true rejections. By borrowing their concepts, the expected number of true rejections for RNA-seq data can be calculated as

$$E(R_1) = \sum_{g\in M_1} \xi(n,\rho_g,\mu_{0g},\phi_g,w,\alpha),$$

where ρ_g is the fold change, ϕ_g is the dispersion, and μ_{0g} is the average read count in the control group for gene $g \in M_1$ (the set of prognostic genes), respectively. Thus, to guarantee an expected number of true rejections, say r_1, and control FDR at a specified level f, we have

$$f = \frac{m_0\alpha}{m_0\alpha + r_1} \quad (3)$$

and

$$r_1 = \sum_{g\in M_1} \xi(n,\rho_g,\mu_{0g},\phi_g,w,\alpha). \quad (4)$$

By solving equation (3) with respect to α, we have

$$\alpha^* = \frac{r_1 f}{m_0(1-f)},$$

where α^* is the marginal type I error level for the expected number of true rejections r_1 at a given FDR f. Replacing α with α^* in (4), we have the function with respect to n as

$$g_1(n) = \sum_{g\in M_1} \xi(n,\rho_g,\mu_{0g},\phi_g,w,\alpha^*) - r_1.$$

Then, by solving $g_1(n) = 0$ via a numerical approach, the required sample size for controlling FDR at level f can be obtained.

To calculate the sample size, we have to estimate all of the fold changes ρ_g, dispersions ϕ_g, and average read counts μ_{0g} of gene g for the set of prognostic genes $g \in M_1$ prior to the RNA-seq experiment. However, we may not have enough information to estimate all of those parameters in practice. To address this issue, we propose the following method to obtain a conservative estimate of the required sample size. Because the power increases as $|\log_2(\rho_g)|$ or μ_{0g} increases and ϕ_g decreases, we suggest

using a common $\rho^* = \arg\min_{g\in M_1}\{|\log_2(\rho_g)|\}$ minimum fold change, $\mu_0^* = \min_{g\in M_1}\{\mu_{0g}\}$ minimum average read count, and $\phi^* = \max_{g\in M_1}\{\phi_g\}$ maximum dispersion to estimate each ρ_g, μ_{0g}, and ϕ_g, respectively. In such cases, it gives a more conservative estimate of the required sample size.

When we use ρ^*, μ_0^*, and ϕ^* to estimate each ρ_g, μ_{0g}, and ϕ_g, $g \in M_1$, in the multiple testing context, α^* and β^* can be calculated as $r_1 f/(m_0(1-f))$ and $1-r_1/m_1$, respectively, where m_1 is the number of prognostic genes. In other words, the power function (2) can be applied in the case of multiple gene comparison, with the replacement of α and β with α^* and β^*.

The procedures for sample size calculation detailed in this section can be summarized as follows:

1. Specify the following parameters:

   m : total number genes for testing;
   m_1 : number of prognostic genes;
   r_1 : number of true rejections;
   f : FDR level;
   w : ratio of normalization factors between two groups;
   $\{\mu_{0g}, g \in M_1\}$: average read counts for prognostic gene g in control group;
   $\{\rho_g, g \in M_1\}$: fold changes for prognostic genes g in control group;
   $\{\phi_g, g \in M_1\}$: dispersion for prognostic genes g in control group;

2. Calculate sample size:

   (a) If all the parameters μ_{0g}, ρ_g, and ϕ_g for each prognostic gene g are known, use a numerical approach to solve the equation below with respect to n.

$$r_1 = \sum_{g\in M_1} \xi(n, \rho_g, \mu_{0g}, \phi_g, w, \alpha^*),$$

   where $\alpha^* = r_1 f/(m_0(1-f))$ and $m_0 = m - m_1$;

   (b) Otherwise,

   (I) specify a desired minimum fold change ρ^*, a minimum average read count μ_0^*, and a maximum dispersion ϕ^*;

   (II) replace $\rho = \rho^*$, $\mu_0 = \mu_0^*$, $\phi = \phi^*$, $\alpha = r_1 f/(m_0(1-f))$, and $\beta = 1 - r_1/m_1$ in equation (2) and solve it with respect to n.

Results

Numerical studies

In this section, we conducted simulation studies to evaluate the accuracy of the proposed sample size formula. The parameter settings in simulation studies are based on empirical data sets.

We set the total number of genes for testing to be $m = 10000$ and the number of statistically significant prognostic genes $m_1 = 100$. We wanted to detect the expected number of true rejections $r_1 = 80$, which corresponds to a power of 80% (i.e. $\beta^* = 0.2$). All parameters μ_{0g}, ρ_g, and ϕ_g ($g = 1, \ldots, 10000$) were assumed to be unknown. Thus, we used a minimum fold change ρ^* and a minimum average read count μ_0^* and a maximum dispersion ϕ^* to estimate each ρ_g, μ_{0g}, and ϕ_g, $g = 1 \ldots, 10000$. We varied $\mu_0^* = 1$ or 5; $\log_2$-fold changes $\log_2(\rho^*) = 0.5, 1.0, 1.5, 2.0$ or 2.5; and $\phi^* = 0.1$, or 0.5. With these settings, $\alpha^* = 8.162 \times 10^{-5}, 4.253 \times 10^{-4}$, and 8.979×10^{-4}, which correspond to controlling FDR at level 1%, 5%, and 10%, respectively.

Then, we substituted α^* and β^* into the formulas (2) and calculated sample size by solving this equation. In addition, for each design setting, we generated 5000 samples from independent negative binomial distributions based on the calculated sample size n; for the control group, the count of each gene is generated by R program from a negative binomial distribution with mean μ_0^* and dispersion ϕ^*; for the treatment group, the count of each gene is generated from a negative binomial distribution with mean $\rho^*\mu_0^*$ and dispersion ϕ^*. Then, edgeR is used to estimate model parameters and perform the exact test. The number of true rejections was counted using the q-value procedure proposed by [20]. The expected number of true rejections was estimated as the sample mean of the number of rejections of the 5000 simulation samples ($\hat{r}_1$).

In Table 1, we showed the calculated sample size with corresponding $\hat{r}_1$ in parentheses under the case $w = 1$. For a fixed $\log_2$-fold change, dispersion, and FDR, sample size increases when μ_0 decreases. This result is as expected; a small average read count provides less information, such that a larger sample size is required to detect the difference. For a fixed μ_0^*, ϕ^*, and FDR, sample size increases when $\log_2(\rho^*)$ decreases (i.e. the smaller $\log_2$-fold changes requires greater sample sizes with all else being equal). This result is as expected; a larger sample size is required for detecting a smaller difference. For a fixed μ_0^*, $\log_2(\rho^*)$, and FDR, sample size increases when ϕ^* increases. This result, also, is as expected; the variation increases when dispersion increases, such that a larger sample size is required to detect the difference. Note that all $\hat{r}_1$ in Table 1 are close to the pre-specified number of true rejections ($r_1 = 80$); thus, the proposed method estimated a sample size that achieves correct power at the specified FDR level.

Table 1 Sample size calculation for simulation study (and $\hat{r}_1$) with $r_1 = 80$ at FDR = 1%, 5% and 10% when $w = 1$, $m = 10000$, $m_1 = 100$

		$\mu_0^* = 1$			$\mu_0^* = 5$		
		FDR			FDR		
$\log_2(\rho^*)$	ϕ^*	1%	5%	10%	1%	5%	10%
0.5	0.1	365 (81)	305 (84)	278 (88)	104 (81)	87 (84)	79 (88)
	0.5	518 (81)	433 (84)	394 (88)	257 (81)	215 (84)	196 (89)
1.0	0.1	79 (81)	67 (84)	61 (87)	24 (82)	20 (84)	19 (91)
	0.5	119 (81)	99 (83)	91 (88)	63 (82)	53 (85)	48 (89)
1.5	0.1	31 (82)	26 (83)	24 (86)	10 (83)	9 (90)	8 (91)
	0.5	49 (81)	41 (83)	38 (88)	28 (83)	23 (84)	21 (86)
2.0	0.1	16 (85)	13 (84)	12 (86)	6 (90)	5 (92)	4 (86)
	0.5	26 (82)	22 (84)	20 (86)	16 (84)	13 (85)	12 (89)
2.5	0.1	8 (85)	7 (89)	6 (87)	3 (78)	3 (81)	3 (98)
	0.5	14 (83)	12 (87)	11 (84)	10 (82)	9 (90)	8 (91)

Applications

Liver and kidney RNA-seq data set

To identify differentially expressed genes between human liver and kidney RNA samples, [8] explored an RNA-seq data set containing 5 human kidney samples and 5 human liver samples. In the following, we used this data set as pilot data for designing a new study with the same study objective. For the purpose of demonstration, we assumed that the human kidney is the control group. After filtering genes with no more than 5 total reads in liver samples or kidney samples, there were 17306 genes left. We assumed that the top 175 ($\approx$ 1% of 17306) genes are prognostic. From the pilot data, the minimum average read counts among the prognostic genes in the control group were estimated as $\mu_0^* = 5$, the maximum dispersion was estimated as $\phi^* = 0.0029$, and the ratio of the geometric mean of normalization factors between the two groups was estimated as $w = 0.9$ using edgeR. Suppose we want to identify 80% of the prognostic genes (i.e. $r_1 = 0.8 \times 175 = 140$), while controlling FDR at 1% (i.e. $f = 0.01$). Based on the pilot data, we set $m = 17306$, $m_1 = 175$, $m_0 = 17131$, $r_1 = 140$, and $f = 0.01$. In this case, we have

$$\alpha^* = \frac{r_1 f}{m_0(1-f)} = 8.2549 \times 10^{-5}$$

and

$$\beta^* = 1 - \frac{r_1}{m_1} = 0.2.$$

After substituting those parameters into equation (2) and solving it with respect to n, the required sample size can be obtained. In the second column from the left of Table 2, we report the sample size while controlling FDR at 1% under various desired minimum fold changes $\rho^* = 0.10, 0.25, 0.50, 0.75, 1.25, 1.50, 2.00, 2.50,$ and 3.0. From Table 2, we found that the original RNA-seq experiment described in [8] with sample size 5 in each group can identify 80% of the prognostic genes at FDR= 1% if the desired minimum fold change ρ^* is 3.0.

Li, 2013 [28] proposed several sample size calculation methods for RNA-seq data under the Poisson model. To compare the difference in sample size calculation between the negative binomial method and Poisson method, in the last six right columns of Table 2 we report the sample size calculation based on Poisson model (i.e. the sample size based on the Wald test n_w, score test n_s, log transformation of Wald statistic n_{lw}, log transformation of score statistic n_{ls}, transformation of Poisson n_{tp}, and likelihood ratio test n_{lr}) with the same settings as the negative binomial method. As we can see, the sample size calculation based on the negative binomial and Poisson methods are similar. This result is as expected since the data set explored by [8] has technical and not biological replicates (i.e. the maximum dispersion estimated from the liver and kidney RNA-seq data set is close to zero). Thus, it is not surprising that the results of the negative binomial and Poisson methods are similar when the dispersion parameter is close to zero. Moreover, in Table 2, the estimated sample size is about the same size for very small fold changes ($\rho^* = 0.10$) and very large fold changes ($\rho^* = 3.0$). This result is expected since it tends to the same conclusion no matter what statistical model is used when the treatment effect is very large (i.e. the fold change is very large or small).

Transcript regulation data set

Blekhman, 2010 [29] used RNA-seq to study transcript regulation in humans, chimpanzees, and rhesus macaques

Table 2 Sample size calculation for liver and kidney RNA-seq data set under various desired minimum fold changes (ρ^*) for $r_1 = 140$ at *FDR* = 1% when $m = 17360$ and $m_1 = 175$

	NB	Poisson					
ρ^*	n	n_w	n_s	n_{lw}	n_{ls}	n_{tp}	n_{lr}
0.10	7	7	7	11	5	5	7
0.25	11	11	11	13	9	9	10
0.50	30	29	30	31	28	27	29
0.75	139	134	136	137	133	132	135
1.25	178	175	173	174	174	177	181
1.50	50	49	48	49	48	50	50
2.00	15	15	15	15	14	16	15
2.50	8	8	8	8	7	8	8
3.00	5	5	5	6	5	6	5

using liver RNA samples from three males and three females from each species. For the purpose of demonstration, we assumed that the goal of the study is to identify differential gene expression between male and female in humans and that the female is considered the control group. There were 13267 genes in the data set after performing quality control analyses. Suppose that the top 133 ($\approx$ 1% of 13267) genes are prognostic. After filtering genes with no more than 5 total reads in male samples or female samples, there were 7658 genes left. Those genes are considered pilot data, and we assessed the differential expression by using edgeR. From the pilot data, the minimum average read counts among the prognostic genes in the control group were estimated as $\mu_0^* = 1.67$, $\phi^* = 0.6513$, and the ratio of the geometric mean of normalization factors between the two groups was estimated as $w = 1.08$. Suppose we want to identify 80% of the prognostic genes (i.e. $r_1 = 0.8 \times 133 = 107$), while controlling the FDR at 10%. Based on the pilot data, we set $m = 13267$, $m_1 = 133$, $m_0 = 13134$, $r_1 = 107$ and $f = 0.1$. In this case, we have $\alpha^* = 9.0512 \times 10^{-4}$ and $\beta^* = 0.2$. In the second column from the left of Table 3, we report the required sample sizes under various desired minimum fold changes while controlling the FDR at 10% under the negative binomial distribution. We also report the required sample size based on the Poisson model proposed by [28] under the same settings in the last six columns on the right of Table 3. As we can see, the required sample size based on the negative binomial method is greater than the Poisson method. In the transcript regulation data set, the maximum dispersion was estimated as $\phi^* = 0.6513 > 0$. This indicates that the read counts in this data set exhibit over-dispersion. In such a situation, it is inappropriate to model this data set based on the Poisson, and the sample size calculation based on the Poisson model will be underestimated due to underestimation of variance (i.e. the study based on the corresponding sample size will be underpowered).

Table 3 Sample size calculation for transcript regulation data set under various desired minimum fold changes (ρ^*) for $r_1 = 107$ at $FDR = 10\%$ when $m = 13267$ and $m_{1=133}$

	NB	Poisson					
ρ^*	n	n_w	n_s	n_{lw}	n_{ls}	n_{tp}	n_{lr}
0.10	19	15	14	21	10	10	14
0.25	35	23	23	26	19	19	21
0.50	109	62	60	62	58	56	59
0.75	558	284	281	282	280	273	281
1.25	821	316	363	366	360	371	381
1.50	240	100	102	103	99	105	105
2.00	79	30	31	32	29	32	32
2.50	44	16	16	18	15	17	16
3.00	30	10	11	12	9	11	10

Discussion

In this research, we assume independent gene expression levels; however, this assumption may not hold in reality. For correlated RNA-seq gene expression data, evaluation of the accuracy of our method is an important future research question; however, generating a negative binomial distribution for correlated high-dimensional data will be a challenge. Moreover, most of the major R packages dedicated to RNA-seq differential analyses (edgeR, DESeq, etc.) are now starting to enable multi-group comparisons. However, the proposed method is developed for comparing two-group means. Thus, the sample size calculation for multi-group comparisons would be an interesting research topic for us in the future. In addition, it has already been noted that typical RNA-seq differential analyses have very low power; see for example the simulation studies in [30], where power for edgeR was always less than 60%, or [31], where power ranged from about 45% to 55% (both with 10 samples per condition). In our simulation and application sections, the minimum sample sizes required to achieve 80% power would be prohibitively large for RNA-seq experiments in practice, given their current cost. In such situations, the findings in [30,31] can provide useful information for specifying achievable power. It is well known that low study power will decrease the reproducibility of scientific research. We hope that this paper can benefit researchers by allowing them to understand their study power.

Conclusions

In recent years, RNA-seq technology has emerged as an attractive alternative to microarray studies, due to its ability to produce digital signals (counts) rather than analog signals (intensities), and to produce more highly reproducible results with relatively little technical variation [32,33]. With a large sample size, RNA-seq can become costly; on the other hand, insufficient sample size may lead to unreliable answers to the research question of interest. To manage the trade-off between cost and accuracy, sample size determination is a critical issue for RNA-seq experimental design. For comparing the differential expression of a single gene, we have proposed a sample size calculation method based on an exact test proposed by [14,15]. To address multiple testing (i.e., multiple genes), we further extended our proposed method to incorporate FDR control. Our methods are not computationally intensive for pilot data or other relevant data with a specified desired minimum fold change, minimum average read count, and maximum

dispersion. To facilitate implementation of the sample size calculation, R code is available from the corresponding author.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Authors CIL, and PFS were involved in the development of the models. CIL and PFS wrote the manuscript. SY generated the original idea and guided and supervised the research. All authors read and approved the final version of this manuscript.

Acknowledgements
This work was partly supported by NIH grants P30CA068485, P50CA095103, P50CA098131, and U01CA163056. The authors wish to thank Margot Bjoring for editorial work on this manuscript.

Author details
[1]Department of Applied Mathematics, National Chiayi University, Chiayi, Taiwan. [2]Department of Statistics, National Cheng Kung University, Tainan, Taiwan. [3]Center for Quantitative Sciences, Vanderbilt University, 571 Preston Building Nashville, TN, USA.

Received: 3 June 2013 Accepted: 28 November 2013
Published: 6 December 2013

References

1. Jiang H, Wong WH: **Statistical inferences for isoform expression in RNA-Seq.** *Bioinformatics* 2009, **25**(8):1026–1032.
2. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty.** *Bioinformatics* 2010, **26**(4):493–500.
3. Wu Z, Wang X, Zhang X: **Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq.** *Bioinformatics* 2011, **27**(4):502–508.
4. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJM, Tai IT, Marra MA: **Alternative expression analysis by RNA sequencing.** *Nat Methods* 2010, **7**(10):843–847.
5. Wang L, Xi Y, Yu J, Dong L, Yen L, Li W: **A statistical method for the detection of alternative splicing using RNA-seq.** *PLoS One* 2010, **5:**e8529.
6. 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.** *Nat Biotechnol* 2010, **28**(5):511–515.
7. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I: **De novo assembly and analysis of RNA-seq data.** *Nat Methods* 2010, **7**(11):909–912.
8. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18**(9):1509–1517.
9. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**(7):613–619.
10. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**(7289):768–772.
11. Auer PL, Doerge RW: **Statistical design and analysis of RNA sequencing data.** *Genetics* 2010, **185**(2):405–416.
12. Fang Z, Cui X: **Design and validation issues in RNA-seq experiments.** *Brief Bioinform* 2011, **12**(3):280–287.
13. Wang L, Feng Z, Wang X, Wang X, Zhang X: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, **26:**136–138.
14. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostat* 2008, **9**(2):321–332.
15. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23**(21):2881–2887.
16. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139–140.
17. Storey JD: **A direct approach to false discovery rates.** *J R Stat Soc Ser B* 2002, **64**(3):479–498.
18. Hirakawa A, Sato Y, Sozu T, Hamada C, Yoshimura I: **Estimating the false discovery rate using mixed normal distribution for identifying differentially expressed genes in microarray data analysis.** *Cancer Inform* 2007, **3:**140–148.
19. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57:**289–300.
20. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci USA* 2003, **100**(16):9440–9445.
21. Pounds S, Cheng C: **Sample size determination for the false discovery rate.** *Bioinformatics* 2005, **21**(23):4263–4271.
22. Hu J, Zou F, Wright FA: **Practical FDR-based sample size calculations in microarray experiment.** *Bioinformatics* 2005, **21:**3264–3272.
23. Jung SH: **Sample size for FDR-control in microarray data analysis.** *Bioinformatics* 2005, **21**(14):3097–3104.
24. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21:**3017–3024.
25. Liu P, Hwang JTG: **Quick calculation for sample size while controlling false discovery rate with application to microarray analysis.** *Bioinformatics* 2007, **23**(6):739–746.
26. Krishnamoorhy K, Thomson J: **A more powerful test for comparing two Poisson means.** *J Stat Plan Infer* 2004, **119:**23–35.
27. Storey JD, Tibshirani R: **Estimating false discovery rates under dependence, with applications to DNA microarrays.** In *Technical Report*. CA: Department of Statistics, Standford University:2001.
28. Li CI, Su PF, Guo Y, Shyr Y: **Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution.** *Int J Comput Biol Drug Des* 2013, **6**(4):358–375.
29. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y: **Sex-specific and lineage-specific alternative splicing in primates.** *Genome Res* 2010, **20**(2):180–189.
30. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14:**91. [http://dx.doi.org/10.1186/1471-2105-14-91]
31. Dillies M, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Brief Bioinform* 2013, **14**(6):671–683.
32. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
33. Hashimoto Si, Qu W, Ahsan B, Ogoshi K, Sasaki A, Nakatani Y, Lee Y, Ogawa M, Ametani A, Suzuki Y, Sugano S, Lee CC, Nutter RC, Morishita S, Matsushima K: **High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer.** *PLoS One* 2009, **4:**e4108.

doi:10.1186/1471-2105-14-357
Cite this article as: Li *et al.*: Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics* 2013 **14**:357.