

# Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution

Chung-I Li, Pei-Fang Su, Yan Guo and Yu Shyr \*

Center for Quantitative Sciences, Vanderbilt University, 571 Preston Building Nashville, TN, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Submitted to Bioinformatics (2011/11/21)

## ABSTRACT

**Motivation:** Sample size determination is an important issue in experimental design of biomedical research. Because of the complexity of RNA-seq experiments, however, the field currently lacks a sample size method widely applicable to differential expression studies utilizing RNA-seq technology.

**Results:** In this report, we propose several methods for sample size calculation for single-gene differential expression analysis of RNA-seq data. These methods then are extended to multiple genes, with consideration for addressing the multiple testing problem by controlling false discovery rate. The methods proposed also can be used to determine optimal sequencing depth; moreover, the methods allow for closed-form sample size formulas with specification of desired minimum fold change and minimum average read count, and thus are not computationally intensive. Simulation studies to evaluate the performance of the proposed sample size formulas are presented; the results indicate our methods work well, with achievement of desired power.

**Availability:** R code for sample size calculation is available from the corresponding author.

**Contact:** yu.shyr@vanderbilt.edu

## 1 INTRODUCTION

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. RNA-seq is more cost-effective than a comparable microarray study; in addition, unlike the microarray chip, which offers only quantification of gene expression level, RNA-seq provides data on expression level as well as differentially spliced variants, gene fusion, and mutation profile. 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 (Jiang and Wong, 2009; Li *et al.*, 2010; Wu *et al.*, 2011), detection of alternative splicing (Griffith *et al.*, 2010; Wang *et al.*, 2010; Trapnell *et al.*, 2010), detection of novel transcripts (Trapnell *et al.*, 2010; Robertson *et al.*, 2010), and the

biology associated with gene expression profile differences between samples (Marioni *et al.*, 2008; Cloonan *et al.*, 2008; Pickrell *et al.*, 2010). 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 (Auer and Doerge, 2010; Fang and Cui, 2011).

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 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 (Marioni *et al.*, 2008; Wang *et al.*, 2010). Further, Fang and Cui (2011) use a Poisson distribution to model RNA-seq data and derive a sample size calculation formula based on the Wald test for single-gene differential expression analysis. However, sample size formulas based on other well-known tests, such as the likelihood ratio test and score test, have not yet been studied, nor have extensions of those test that may improve performance. Therefore, the first goal of this paper is to derive sample size formulas based on the likelihood ratio test, score test, and extensions of those tests, and compare their performance.

In general, sample size is determined as part of the experimental design. Due to limited funding or difficulties obtaining sample, however, sample size often is pre-determined, prior to the experimental design stage. In such cases, it is critical to decide optimal sequencing depth for each sample to attain a desired power. As expected, power to detect differential expression increases as sequencing depth increases; over-sequencing (too much depth) may lead to a waste of resource without significant increase in power, while under-sequencing (too little depth) may result in insufficient power to detect true differences. Thus, when sample size is predetermined, finding the optimal trade-off between cost (depth) of sequencing and statistical power is also a critical issue for design of an RNA-seq experiment. To our knowledge, determination of optimal sequencing depth is not yet addressed in the literature; thus, our second question is: what sequencing depth is required to attain a specified power for identifying differential gene expression between two groups?

In reality, thousands of genes are examined in an RNA-seq experiment; differential expression among those genes is tested

\*To whom correspondence should be addressed

simultaneously, requiring correction of error rates for multiple comparisons. Several such corrected measures have been proposed, such as family-wise error rate (FWER) and false discovery rate (FDR). In the multiple testing circumstance, controlling FDR is preferable (Storey, 2002) because the Bonferroni correction for FWER is often too conservative (Hirakawa *et al.*, 2007). Many methods have been proposed to control FDR in the analysis of high-dimensional data (Benjamini and Hochberg, 1995; Storey, 2002; Storey and Tibshirani, 2003). Those concepts have been extended to calculate sample size for microarray studies (Pounds and Cheng, 2005; Hu *et al.*, 2005; Jung, 2005; Pawitan *et al.*, 2005; Liu and Hwang, 2007). To our knowledge, however, the literature does not address determination of sample size while controlling FDR in RNA-seq data. Therefore, the third 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 questions. (i) For single gene testing, what is the optimal sample size to attain a specified power for identifying differential gene expression between two groups? (ii) When sample size is fixed, what sequencing depth is required to attain a specified power for identifying differential gene expression between two groups? (iii) For multiple gene comparison, what is the suitable sample size while controlling FDR? The article is organized as follows. In Section 2, we describe several test statistics for identifying differential expression between two groups; the corresponding sample size and depth calculation formulas are proposed for comparing a single gene. In Section 3, we extend those methods to address the multiple testing problem. Performance comparisons via numerical studies are described in Section 4. Finally, discussion follows in Section 5.

## 2 METHODS

### 2.1 Notations

In this section, we focus on single-gene comparison. Let  $X_{ij}$  be the observed count of mapping reads in the  $j$ th ( $j = 1, 2, \dots, n_i$ ) sample of the  $i$ th ( $i = 0, 1$ ) group, where  $n_0$  and  $n_1$  are the numbers of samples from the control and treatment group, respectively. Assume read counts within each group are independent.  $X_{ij}$  can be modeled as a Poisson random variable with parameter  $d_{ij}\gamma_i$ , where  $\gamma_i$  represents the gene expression level of group  $i$  and  $d_{ij}$  represents the total number of reads mapped in the  $j$ th ( $j = 1, 2, \dots, n_i$ ) sample of the  $i$ th ( $i = 0, 1$ ) group. The question of interest is the identification of differential gene expression between two groups; the corresponding testing hypothesis is

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

Because the problem of determining an adequate sample size can be addressed through the traditional hypothesis testing framework, we extend and modify several test statistics for RNA-seq data in the following section.

### 2.2 Test statistics

For an RNA-seq experiment, we can easily count the total number of reads for the control group  $X_0$  and treatment group  $X_1$ . Because  $X_{ij}$  is a Poisson random variable with parameter  $d_{ij}\gamma_i$ ,  $X_i = \sum_{j=1}^{n_i} X_{ij}$  follows a Poisson distribution with parameter  $\mu_i = d_i\gamma_i$ ,

where  $d_i = \sum_{j=1}^{n_i} d_{ij}$  is the total number of mapping reads in group  $i$ , and  $\mu_i$  can be treated as the average read count of condition  $i$ . Then, to compare two independent Poisson variables with unequal frames, we apply the method proposed by Ng and Tang (2005) for testing the null hypothesis (1). Following the works of Ng and Tang (2005), we have the Wald ( $Z_w$ ) and Score ( $Z_s$ ) statistics

$$Z_w = \frac{X_1 - wX_0}{\sqrt{X_1 + w^2X_0}}, \quad (2)$$

$$Z_s = \frac{X_1 - wX_0}{\sqrt{(X_1 + X_0)w}}, \quad (3)$$

respectively, where  $w = d_1/d_0$  is the ratio of total number of reads mapped between two groups.

Logarithmic transformation is usually adopted for skewness correction and variance stabilization (Ng and Tang, 2005). Accordingly, the log-transformations of (2) and (3) are

$$Z_{lw} = \frac{\ln(X_1/X_0) - \ln w}{\sqrt{1/X_1 + 1/X_0}}, \quad (4)$$

$$Z_{ls} = \frac{\ln(X_1/X_0) - \ln w}{\sqrt{(2 + w + 1/w)/(X_1 + X_0)}}. \quad (5)$$

In addition, based on the concept of transformation of a Poisson random variable suggested by Huffman (1984) and Gu *et al.* (2008), we have

$$Z_{tp} = \frac{2(\sqrt{X_1 + 3/8} - \sqrt{w(X_0 + 3/8)})}{\sqrt{1 + w}}, \quad (6)$$

which can accelerate the rate of convergence to normality. Last, we consider another method, the likelihood ratio test (LRT), and we derive the exact form of the LRT statistic as

$$Z_{lr} = \left\{ \left( \frac{X_1 + X_0}{X_1(1 + 1/w)} \right)^{X_1} \left( \frac{X_1 + X_0}{X_0(1 + w)} \right)^{X_0} \right\}. \quad (7)$$

This is the ratio of the maximum value of the likelihood function under the constraint of the null hypothesis to the maximum value of the likelihood function with the constraint relaxed. The above statistics, (2)-(7), can be used for testing the hypothesis (1).

Because statistics (2)-(6) are asymptotically standard normal distribution under the null hypothesis, the approximate  $p$ -values of the corresponding test statistics are calculated as

$$2(1 - \Phi(|Z(x_1, x_0)|)), \quad (8)$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution, and  $Z(x_1, x_0)$  is any one of the observed values of (2) to (6). In addition, for LRT, since  $-2 \ln(Z_{lr})$  approximately follows a chi-square distribution with one degree of freedom, the approximate  $p$ -value is given by

$$1 - \chi_1^2(Z_{lr}(x_1, x_0)), \quad (9)$$

where  $Z_{lr}(x_1, x_0)$  is the observed value in (7) and  $\chi_1^2$  is the cumulative distribution function of the chi-square distribution with one degree of freedom. For a given level of significance  $\alpha$ , we reject  $H_0$  in (1) when the  $p$ -value from (8) or (9) is less than  $\alpha$ .

### 2.3 Sample size calculation

In this section, we focus on sample size calculation based on the test statistics described in Section 2.2. For simplicity, we assume the RNA-seq experiment uses a balanced design (i.e.,  $n_1 = n_2 = n$ ), which is a special but common case. For attaining a specified power  $1 - \beta$  and significance level  $\alpha$  to detect the fold change of interest  $\rho = \gamma_1/\gamma_0$ , we derive the sample size formula for each test statistic as summarized below; details of the derivation are provided in the supplementary materials. For the given parameters, the sample size formula based on the Wald test ( $n_w$ ), score test ( $n_s$ ), log transformation of Wald statistic ( $n_{lw}$ ), log transformation of score statistic ( $n_{ls}$ ), and transformation of Poisson ( $n_{tp}$ ) are

$$n_w = \frac{(1 + \rho/w)(z_{1-\alpha/2} + z_{1-\beta})^2}{\mu_0(\rho - 1)^2}, \quad (10)$$

$$n_s = \frac{(1 + \rho/w)(z_{1-\alpha/2} \sqrt{\frac{1+w\rho}{w+\rho}} + z_{1-\beta})^2}{\mu_0(\rho - 1)^2}, \quad (11)$$

$$n_{lw} = \frac{(1 + \frac{1}{\rho w})(z_{1-\alpha/2} + z_{1-\beta})^2}{\mu_0(\ln \rho)^2}, \quad (12)$$

$$n_{ls} = \frac{(1 + \frac{1}{\rho w})(z_{1-\alpha/2} \sqrt{\frac{2+w+1/w}{(1+w\rho)(1+1/w\rho)}} + z_{1-\beta})^2}{\mu_0(\ln \rho)^2}, \quad (13)$$

$$n_{tp} = \frac{\left( \frac{z_{1-\alpha/2} \sqrt{(1+w)/w} + z_{1-\beta} \sqrt{(1+w\rho)/w}}{2(\sqrt{\rho}-1)} \right)^2 - \frac{3}{8}}{\mu_0}, \quad (14)$$

respectively, where  $z_{1-\alpha/2}$  is the 100(1- $\alpha$ /2) percentile of the standard normal distribution,  $\mu_0 = \bar{d}_0\gamma_0$  is the average read count in the control group,  $\bar{d}_0$  is the average sequencing depth in the control group, and  $\gamma_0$  is control group gene expression level. In practice, for detecting a desired fold change  $\rho$ , we first must specify significance level  $\alpha$  and power  $1 - \beta$ . Then, the ratio of total number of reads between the two groups  $w$  and the average read count in the control group  $\mu_0$  can be estimated from pilot data or other relevant studies.

For LRT, a closed form to calculate sample size is difficult to derive. Krishnamoorthy and Thomson (2004) have shown the power of LRT can be expressed as a function of sample size in the form

$$1 - \beta = \sum_{x_0=0}^{\infty} \sum_{x_1=0}^{\infty} \frac{e^{-n\mu_0} (n\mu_0)^{x_0} e^{-n\rho\mu_0} (n\rho\mu_0)^{x_1}}{x_0!x_1!} I(p < \alpha), \quad (15)$$

where  $p$  is the  $p$ -value calculated by using (9) and  $I(\cdot)$  denotes the indicator function. Required sample size  $n_{lr}$  then may be calculated by solving (15) through a numerical approach, such as a gradient-search or bisection procedure.

### 2.4 Sequencing depth determination

Here we are interested in determining optimal sequencing depth to achieve a specified power for testing hypothesis (1), when the sample size  $n$  is predetermined. To address this issue, we can substitute  $\mu_0 = \bar{d}_0\gamma_0$  into (10)-(14). Then, the required average

sequencing depth for a specified power under each method can be derived as

$$\bar{d}_w = \frac{(1 + \rho/w)(z_{1-\alpha/2} + z_{1-\beta})^2}{n\gamma_0(\rho - 1)^2}, \quad (16)$$

$$\bar{d}_s = \frac{(1 + \rho/w)(z_{1-\alpha/2} \sqrt{\frac{1+w\rho}{w+\rho}} + z_{1-\beta})^2}{n\gamma_0(\rho - 1)^2}, \quad (17)$$

$$\bar{d}_{lw} = \frac{(1 + \frac{1}{\rho w})(z_{1-\alpha/2} + z_{1-\beta})^2}{n\gamma_0(\ln \rho)^2}, \quad (18)$$

$$\bar{d}_{ls} = \frac{(1 + \frac{1}{\rho w})(z_{1-\alpha/2} \sqrt{\frac{2+w+1/w}{(1+w\rho)(1+1/w\rho)}} + z_{1-\beta})^2}{n\gamma_0(\ln \rho)^2}, \quad (19)$$

$$\bar{d}_{tp} = \frac{\left( \frac{z_{1-\alpha/2} \sqrt{(1+w)/w} + z_{1-\beta} \sqrt{(1+w\rho)/w}}{2(\sqrt{\rho}-1)} \right)^2 - \frac{3}{8}}{n\gamma_0}, \quad (20)$$

respectively.

Similarly, the average sequencing depth required for LRT can be computed by solving  $\bar{d}$  in equation

$$1 - \beta = \sum_{x_0=0}^{\infty} \sum_{x_1=0}^{\infty} \frac{e^{-n\bar{d}\gamma_0} (n\bar{d}\gamma_0)^{x_0} e^{-n\bar{d}\rho\gamma_0} (n\bar{d}\rho\gamma_0)^{x_1}}{x_0!x_1!} I(p < \alpha) \quad (21)$$

by substituting  $\mu_0 = \bar{d}_0\gamma_0$  into (15).

## 3 SAMPLE SIZE DETERMINATION FOR 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 this situation, the sample size formulas discussed above cannot be applied directly. In this section, we extend the sample size formulas thus far derived to address this issue.

To address the multiple testing problem, Benjamini and Hochberg (1995) suggested the use of false discovery rate (FDR) rather than type I error rate. FDR is defined as expected proportion of false discoveries among rejected null hypotheses. Storey (2002) further proposed an improvement to FDR to achieve higher power, in the form

$$\text{FDR} = E\left(\frac{R_0}{R} \mid R > 0\right), \quad (22)$$

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).

For sample size calculation, Jung (2005) proposed an FDR-controlled method for microarray data analysis, based on the expression of FDR under independence (or weak dependence) among test statistics, as

$$\text{FDR} = \frac{m_0\alpha}{m_0\alpha + E(R_1)}, \quad (23)$$

(Storey, 2002; Storey and Tibshirani, 2001), where  $m_0$  is the number of true null hypotheses, and  $E(R_1)$  is the expected number

of true rejections.  $E(R_1)$  can be calculated as

$$E(R_1) = \sum_{j \in M_1} \xi_j(\alpha), \quad (24)$$

where  $\xi_j(\cdot)$  is the power function of the single  $\alpha$ -level test for gene  $j \in M_1$  (the set of prognostic genes). To derive the sample size formula for RNA-seq data, we modified the power function in (24) based on the test statistics in (2)-(7), as follows. That is, to guarantee an expected number of true rejections, say  $r_1$ , and control FDR at a specified level  $f$ , we can rewrite (23) and (24) as

$$f = \frac{m_0 \alpha}{m_0 \alpha + r_1}, \quad (25)$$

and

$$r_1 = \sum_{j \in M_1} \xi(\rho_j, \mu_{0j}, \alpha, w, n). \quad (26)$$

By solving equation (25) with respect to  $\alpha$ , we have

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

where  $\alpha^*$  is the marginal type I error level for the expected number of true rejections  $r_1$  at a given FDR  $f$ . Replacing  $\alpha$  with  $\alpha^*$  in (26), we have the function with respect to  $n$  as

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

Then, by solving  $g_1(n) = 0$  via a numerical approach (i.e., gradient-search or bisection procedure), the sample size for controlling FDR at level  $f$  can be obtained.

In practice, among the set of prognostic genes, we may not have enough information to estimate each fold change  $\rho_j$  and average read count  $\mu_{0j}$  prior to the RNA-seq experiment. Therefore, we suggest using a common minimum fold change  $\rho^* = \arg \min_{j \in M_1} \{|\rho_j - 1|\}$ , and minimum average read count  $\mu_0^* = \min_{j \in M_1} \{\mu_{0j}\}$ , to estimate each  $\rho_j$  and  $\mu_{0j}$ , respectively.

When we use  $\rho^*$  and  $\mu_0^*$  in equation  $g_1(n)$  to estimate each  $\rho_j$  and  $\mu_{0j}$ ,  $j \in M_1$ , we are able to generate easy-to-use, closed-form sample size formulas, corresponding to (10)-(14). In the multiple testing context,  $\alpha^*$  and  $\beta^*$  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 sample size formulas (10)-(15) derived in Section 2.3 can be applied in the multiple-testing case, with the replacement of  $\alpha$  and  $\beta$  with  $\alpha^*$  and  $\beta^*$ . This reduces to the conventional sample size calculation when we want to detect a fold change of  $\rho^*$  with power  $1 - \beta^* = r_1 / m_1$  while controlling type I error at  $\alpha^* = r_1 f / (m_0(1-f))$ .

The procedures for sample size calculation detailed in the sections above may 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 total number of reads between two groups;

$\{\mu_{0j}, j \in M_1\}$ : average read counts in control group;

$\{\rho_j, j \in M_1\}$ : fold changes for prognostic genes;

2. Calculate sample size:
  - a. If all the parameters  $\mu_{0j}$  and  $\rho_j$  for each gene are known, use a numerical approach to solve equation

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

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

- b. Otherwise,
  - (I) specify a desired minimum fold change  $\rho^*$  and a minimum average read count  $\mu_0^*$ ;
  - (II) replace  $\rho = \rho^*$ ,  $\mu_0 = \mu_0^*$ ,  $\alpha = r_1 f / (m_0(1-f))$ , and  $\beta = 1 - r_1 / m_1$  in equations (10)-(15).

## 4 NUMERICAL STUDIES

The proposed sample size formulas are based on the large sample theory and a nearby alternative hypothesis approximation. To evaluate the performance of the formula under a finite sample and fixed alternative hypothesis, we conducted simulations. In the first simulation, we focus on the single-gene testing problem and calculate required sample size. In the second simulation, we integrate the FDR issue into the sample size procedure.

### 4.1 Simulation study of sample size calculation for single-gene testing

First, we focus on the single-gene testing problem. We assume the following input settings: type I error rate  $\alpha = 0.05$ ; power  $1 - \beta = 0.8$ ;  $w = 0.5, 1.0$ , or  $2.0$ ;  $\mu_0 = 1, 5$ , or  $10$ ; and  $\rho = 1.25, 0.75, 1.5$  or  $0.5$  (i.e., the corresponding absolute  $\log_2$ -fold changes ( $|\log_2(\rho)|$ ) are  $0.32, 0.42, 0.58$  and  $1.00$ ). For each combination of design settings, we calculate the required sample size using our derived formulas, and then 5000 simulation samples of size  $n$  are generated from the same setting. The test statistics in (2)-(7) also are applied to each simulation sample, and the empirical power is obtained as the proportion of simulation samples for which  $H_0$  is rejected with  $\alpha = 0.05$ .

Table 1 reports the sample sizes calculated from (10)-(15), with associated empirical power given in parentheses. For a fixed fold change, sample size decreases when  $\mu_0$  increases. This result is as expected; for a fixed fold change, small average read count provides less information, such that a larger sample size is required to detect the difference. Similarly, for a fixed  $\mu_0$ , sample size decreases when  $|\log_2(\rho)|$  increases. This result, also, is as expected; a larger sample size is required for detecting a smaller difference. Under all design settings, all empirical power corresponding to  $n$  are close to nominal power  $0.8$ , indicating our methods achieve power at the desired level. With regard to  $n$ , no one method consistently performs better than the others; results from the simulation study may be summarized as follows: For  $w < 1$  and  $\rho < 1$ ,  $n_w$  is smaller than others; for  $w < 1$  and  $\rho > 1$ ,  $n_{1w}$  is smaller than others; for  $w = 1$  and  $\rho < 1$ ,  $n_{tp}$  is smaller than others; for  $w = 1$  and  $\rho > 1$ ,  $n_{ls}$  is smaller than others; for  $w > 1$  and  $\rho < 1$ ,  $n_{tp}$  is smaller than others; and for  $w > 1$  and  $\rho > 1$ ,  $n_w$  is smaller than others.

**Table 1.** Sample size (and empirical power) under each design setting with  $\alpha = 0.05$  and  $1 - \beta = 0.8$

$w$	$\rho$ ( $ \log_2(\rho) $ )	$\mu_0$	$n_w$	$n_s$	$n_{lw}$	$n_{ls}$	$n_{tp}$	$n_{lr}$
0.5	1.25 (0.32)	1	440 (.81)	418 (.80)	410 (.80)	429 (.80)	433 (.80)	442 (.81)
		5	88 (.80)	84 (.80)	82 (.80)	86 (.81)	87 (.81)	89 (.81)
		10	44 (.81)	42 (.80)	41 (.80)	43 (.80)	44 (.82)	45 (.81)
	0.75 (0.42)	1	314 (.80)	336 (.80)	348 (.81)	322 (.81)	320 (.79)	329 (.81)
		5	63 (.80)	68 (.81)	70 (.81)	65 (.81)	64 (.80)	66 (.80)
		10	32 (.81)	34 (.81)	35 (.81)	33 (.81)	32 (.79)	33 (.80)
	1.5 (0.58)	1	126 (.81)	115 (.81)	112 (.80)	120 (.80)	122 (.82)	122 (.82)
		5	26 (.82)	23 (.80)	23 (.81)	24 (.80)	25 (.82)	25 (.83)
		10	13 (.82)	12 (.81)	12 (.82)	12 (.81)	13 (.84)	13 (.85)
	0.5 (1.00)	1	63 (.80)	74 (.81)	82 (.86)	66 (.81)	70 (.81)	69 (.80)
		5	13 (.80)	15 (.82)	17 (.86)	14 (.84)	13 (.79)	14 (.80)
		10	7 (.83)	8 (.80)	9 (.88)	7 (.84)	7 (.80)	7 (.81)
1.0	1.25 (0.32)	1	283 (.80)	283 (.80)	284 (.80)	282 (.80)	292 (.81)	304 (.83)
		5	57 (.80)	57 (.81)	57 (.81)	57 (.81)	59 (.81)	61 (.82)
		10	29 (.81)	29 (.81)	29 (.82)	29 (.81)	30 (.82)	31 (.83)
	0.75 (0.42)	1	220 (.80)	220 (.80)	222 (.81)	219 (.80)	210 (.79)	219 (.80)
		5	44 (.81)	44 (.80)	45 (.81)	44 (.81)	42 (.79)	44 (.80)
		10	22 (.80)	22 (.80)	23 (.81)	22 (.80)	21 (.80)	22 (.80)
	1.5 (0.58)	1	79 (.80)	79 (.81)	80 (.81)	78 (.80)	83 (.82)	82 (.82)
		5	16 (.81)	16 (.80)	16 (.81)	16 (.81)	17 (.83)	17 (.83)
		10	8 (.81)	8 (.81)	8 (.81)	8 (.81)	9 (.86)	9 (.85)
	0.5 (1.00)	1	48 (.81)	48 (.82)	50 (.82)	46 (.80)	42 (.76)	46 (.81)
		5	10 (.83)	10 (.84)	10 (.84)	10 (.84)	9 (.80)	10 (.83)
		10	5 (.82)	5 (.85)	5 (.85)	5 (.83)	5 (.83)	5 (.83)
2.0	1.25 (0.32)	1	205 (.80)	215 (.81)	221 (.80)	208 (.80)	222 (.82)	229 (.82)
		5	41 (.80)	43 (.81)	45 (.81)	42 (.81)	45 (.82)	46 (.83)
		10	21 (.81)	22 (.82)	23 (.83)	21 (.80)	23 (.83)	23 (.83)
	0.75 (0.42)	1	173 (.81)	162 (.80)	159 (.80)	167 (.80)	156 (.79)	164 (.80)
		5	35 (.81)	33 (.80)	32 (.80)	34 (.80)	32 (.80)	33 (.80)
		10	18 (.83)	17 (.82)	16 (.80)	17 (.80)	16 (.80)	17 (.81)
	1.5 (0.58)	1	55 (.80)	61 (.81)	64 (.82)	57 (.81)	64 (.83)	64 (.83)
		5	11 (.80)	13 (.83)	13 (.83)	12 (.83)	13 (.85)	13 (.84)
		10	6 (.82)	7 (.85)	7 (.85)	6 (.83)	7 (.85)	7 (.86)
	0.5 (1.00)	1	40 (.83)	34 (.81)	33 (.80)	36 (.80)	31 (.78)	35 (.81)
		5	8 (.85)	7 (.82)	7 (.81)	8 (.84)	7 (.81)	7 (.81)
		10	4 (.83)	4 (.87)	4 (.86)	4 (.86)	4 (.86)	4 (.86)

**Table 2.** Sample size (and  $\hat{r}_1$ ) for  $r_1 = 32$  at FDR = 0.05 when  $w = 1$ ,  $m = 4000$ ,  $m_1 = 40$

$ \log_2(\rho) $	$\mu_0$	$n_w$	$n_s$	$n_{lw}$	$n_{ls}$	$n_{tp}$	$n_{lr}$
0.32	1	687 (32)	687 (32)	689 (33)	683 (32)	700 (33)	720 (32)
	5	138 (32)	138 (32)	138 (33)	137 (32)	140 (33)	144 (32)
	10	69 (32)	69 (32)	69 (33)	69 (33)	70 (33)	72 (32)
0.42	1	534 (32)	534 (32)	538 (33)	529 (32)	518 (32)	532 (32)
	5	107 (32)	107 (32)	108 (33)	106 (32)	104 (32)	107 (32)
	10	54 (33)	54 (33)	54 (33)	53 (32)	52 (32)	54 (33)
0.58	1	191 (32)	191 (32)	194 (33)	187 (32)	197 (33)	198 (32)
	5	39 (33)	39 (33)	39 (33)	38 (32)	40 (34)	40 (34)
	10	20 (34)	20 (34)	20 (33)	19 (33)	20 (34)	20 (34)
1.00	1	115 (33)	115 (33)	119 (33)	109 (33)	106 (31)	112 (33)
	5	23 (33)	23 (33)	24 (34)	22 (33)	22 (32)	23 (33)
	10	12 (34)	12 (34)	12 (34)	11 (34)	11 (32)	12 (34)
1.50	1	22 (33)	22 (34)	24 (35)	20 (33)	24 (36)	22 (33)
	5	5 (36)	5 (36)	5 (36)	4 (33)	5 (37)	5 (37)
	10	3 (39)	3 (39)	3 (38)	2 (34)	3 (39)	3 (39)

**Table 3.** Guidelines for implementing sample size calculation

	$\rho < 1$	$\rho > 1$
$w < 1$	$n_w$	$n_{lw}$
$w = 1$	$n_{tp}$	$n_{ls}$
$w > 1$	$n_{tp}$	$n_w$

As seen in Table 2, all  $\hat{r}_1$  are close to the pre-specified number of true rejections ( $r_1 = 32$ ); thus, the proposed formulas estimate a sample size that achieves correct power at the specified FDR level. Under the same design settings, the sample sizes in Table 2 are larger than those in Table 1; as expected, we need a larger sample size for testing differential expression among thousands of genes simultaneously than for testing a single gene. With regard to comparison of methods within Table 2, we observed patterns similar to those described for Table 1, with no one method better than the others in all settings. In particular settings, however, each method tends to show consistent superior performance, providing the minimum required sample size among all methods in that setting. Therefore, we provide recommendations in Table 3, based on the observations from our simulation studies, to facilitate choice of sample size calculation to achieve minimum required sample size.

## 4.2 Simulation study of sample size calculation for multiple gene testing

In the second simulation study, we evaluate performance of the sample size formulas while controlling FDR. We set the total number of genes for testing  $m = 4000$ , and number of prognostic genes  $m_1 = 40$ . We wish to detect the expected number of true rejections  $r_1 = 32$ , which corresponds to a power of 80%. All parameters  $\mu_j$  and  $\rho_j$  ( $j = 1, \dots, 4000$ ) are assumed to be unknown. Thus, we use a minimum fold change  $\rho^*$  and a minimum average read count  $\mu^*$  to estimate each  $\rho_j$  and  $\mu_j$ ,  $j = 1, \dots, 4000$ . We vary  $\mu^* = 1, 5$ , or 10;  $w = 0.5, 1.0$ , or 2.0; and  $\rho^* = 1.25, 0.75, 1.5, 0.5$  or 2.83 (i.e., the corresponding absolute  $\log_2$ -fold changes ( $|\log_2(\rho)|$ ) are 0.32, 0.42, 0.58, 1.00 and 1.50, respectively). In these settings,  $\alpha^* = 4.253 \times 10^{-4}$  and  $\beta^* = 0.2$  are calculated for controlling FDR at level 0.05. Then, we can substitute  $\alpha^*$  and  $\beta^*$  in the formulas (10)-(15) to calculate sample size under each method. In addition, for each design setting, we generate 5000 samples from independent Poisson distributions based on the calculated sample size  $n$ . The number of true rejections is counted using the q-value procedure proposed by Storey and Tibshirani (2003). The expected number of true rejections is estimated as the sample mean of the 5000 simulation samples ( $\hat{r}_1$ ). In Table 2, we show calculated sample size with corresponding  $\hat{r}_1$  in parentheses under the case  $w = 1$ . The results for  $w = 0.5$  or 2.0 are shown in the supplementary materials.

## 5 CONCLUSION

In recent years, RNA-seq technology has emerged as an attractive alternative to microarray studies, due to the ability to produce digital signals (counts) rather than analog signals (intensities), and to produce more highly reproducible results with relatively little technical variation (Mortazavi *et al.*, 2008; Hashimoto *et al.*, 2009). 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 differential expression of a single gene, we have proposed sample size calculation formulas based on several test statistics. To address

multiple testing (i.e., multiple genes), we further extend our proposed sample size formulas to incorporate FDR control. Our methods are not computationally intensive, with the exception of the numerical approach required to estimate sample size under the LRT method; with the exception of this method, our proposed sample size formulas have easy-to-use, closed forms, given pilot or other relevant data for specification of a desired minimum fold change and minimum average read count. In addition, to determine optimal sequencing depth when sample size is fixed, we have derived sequencing depth determination formulas as described in Section 2.4. To facilitate implementation of sample size calculation or sequencing depth determination, R code is available from the corresponding author.

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 accuracy of our methods is an important future research question; however, generation of a Poisson distribution for correlated high-dimensional data will be a challenge. Another important extension to our research involves the count data assumption; RNA-seq data are count data, and we assume the read counts follow a Poisson distribution naturally. A critical assumption of the Poisson model is the mean and variance are equal; however, this assumption may not hold, as count of reads could exhibit variation significantly greater than the mean (Robinson and Smyth, 2008). To deal with over-dispersion, several approaches for testing differential gene expression in RNA-seq data have been proposed (e.g., negative binomial model, Anders and Huber, 2010; generalized Poisson model, Srivastava and Chen, 2010). Therefore, extending our sample size methods to address over-dispersion will allow for wider applicability of these methods.

## ACKNOWLEDGEMENT

The authors wish to thank Lynne Berry for editorial work on this manuscript.

*Funding:* This work was supported by NIH research grants P30 CA068485, P50 CA090949, P50 CA095103, and P50 CA098131.

## REFERENCES

- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Auer, P.L. and Doerge, R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405-416.
- Benjamini, Y. and Hochberg, Y. (1995) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843-847.
- Cloonan, N. et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613-619.
- Fang, Z. and Cui, X. (2011) Design and validation issues in RNA-seq experiments. *Brief. Bioinform.*, **12**, 280-287.
- Griffith, M. et al. (2010) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289-300.
- Gu, K. et al. (2008) Testing the ratio of two Poisson rates. *Biom. J.*, **50**, 283-298.
- Hashimoto, S.I. et al. (2009) High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PLoS One*, **4**, e4108.
- Hirakawa, A. et al. (2007) Estimating the false discovery rate using mixed normal distribution for identifying differentially expressed genes in microarray data analysis. *Cancer Inform.*, **3**, 140-148.
- Hu, J. et al. (2005) Practical FDR-based sample size calculation in microarray experiments. *Bioinformatics*, **21**, 3264-3272.
- Huffman, M.D. (1984) An improved approximate two-sample Poisson test. *Appl. Statist.*, **33**, 224-226.
- Jiang, H. and Wong, W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032.
- Jung, S.H. (2005) Sample size for FDR-control in microarray data analysis. *Bioinformatics*, **21**, 3097-3104.
- Krishnamoorthy, K. and Thomson, J. (2004) A more powerful test for comparing two Poisson means. *J. Stat. Plan Infer.*, **119**, 23-35.
- Li, B. et al. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493-500.
- Liu, P. and Hwang, J.T. (2007) Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, **21**, 3097-3104.
- Marioni, J.C. et al. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509-1517.
- Mortazavi, A. et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628.
- Ng, H.K.T. and Tang, M.L. (2005) Testing the equality of two Poisson means using the rate ratio. *Stat. Med.*, **24**, 955-965.
- Pawitan, Y. et al. (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017-3024.
- Pickrell, J.K. et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768-772.
- Pounds, S. and Cheng, C. (2005) Sample size determination for the false discovery rate. *Bioinformatics*, **21**, 4263-4271.
- Robinson, M.D. and Smyth, G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **8**, 321-332.
- Robertson, G. et al. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909-912.
- Srivastava, S. and Chen, L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.
- Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B*, **64**, 479-498.
- Storey, J.D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report 2001-28*, Department of Statistics, Stanford University, CA.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci.*, **100**, 9440-9445.
- Trapnell, C. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511-515.
- Wang, L. et al. (2010) A statistical method for the detection of alternative splicing using RNA-seq. *PLoS One*, **5**, e8529.
- Wang, L. et al. (2010) DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136-138.

Wu,Z. *et al.* (2011) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq.

*Bioinformatics*, **27**, 502-508.

Technical Report  
Submitted to Bioinformatics  
(2011/11/21)