Published in final edited form as:

Int J Comput Biol Drug Des. 2013 ; 6(4): . doi:10.1504/IJCBDD.2013.056830.

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

Chung-I Li: chung-i.li@vanderbilt.edu; Pei-Fang Su: pei-fang.su@vanderbilt.edu; Yan Guo: yan.guo@vanderbilt.edu; Yu Shyr: yu.shyr@vanderbilt.edu

Abstract

Sample size determination is an important issue in the 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. In this report, we propose several methods for sample size calculation for single-gene differential expression analysis of RNA-seq data under Poisson distribution. These methods are then extended to multiple genes, with consideration for addressing the multiple testing problem by controlling false discovery rate. Moreover, most of the proposed methods allow for closed-form sample size formulas with specification of the 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 that our methods work well, with achievement of desired power. Finally, our sample size calculation methods are applied to three real RNA-seq data sets.

Keywords

sample size determination; RNA-seq; false discovery rate; Poisson distribution

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 microarray study, and 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., 2010a; 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 the 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 in order to identify differential gene expression (Marioni et al., 2008; Wang et al., 2010b). Further, Fang and Cui (2011) used 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 tests 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 reality, thousands of genes are examined in an RNA-seq experiment; differential expression among those genes is tested 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 multiple testing circumstances, 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). The sample size calculation methods for microarray studies are developed based on two-sample t-test under Gaussian distribution assumption. However, because RNA-seq count data often have skewed distributions, the t-test is inappropriate for count data. Thus, the sample size calculation method for microarray studies cannot be applied to RNA-seq count data directly. 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 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) 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 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. A real RNA-seq data set is used to illustrate sample size calculation in Section 5. Finally, discussion follows in Section 6.

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 ($j = 1, 2, \dots, n_j$) 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 groups, 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 γ_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. Because the problem of determining an adequate sample size can be addressed through the traditional hypothesis testing framework, we describe several test statistics for RNA-seq data in the following section.

2.2 Test statistics

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 μ_i can be treated as the average read count of condition i . For identifying differential gene expression between two groups, the corresponding testing hypothesis is

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

The testing hypothesis described above is equivalent to comparing two independent Poisson rates with unequal sampling frames. To address this issue, we follow the works of Ng and Tang (2005), and 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 the total number of reads mapped between the two groups. In most of cases, the total numbers of reads mapped between the two groups are not equal (i.e. $w \neq 1$). 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

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

which can accelerate the rate of convergence to normality. Finally, 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. In most of the cases, the total numbers of reads mapped between two groups are not equal (i.e. $w \neq 1$). The statistics (2)–(7) can take that information into account, and they can be used for testing the hypothesis (1).

Because statistics in (2)–(6) have an 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 the LRT in (7), 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 χ_1^2 is the cumulative distribution function of the chi-square distribution with one degree of freedom. For a given level of significance α , we reject H_0 in (1) when the p -value from (8) or (9) is less than α .

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. The following method could be easily extended to the unbalanced case (i.e. let $n_0 = n$ and $n_1 = k_n$ where k is a predetermined ratio of the sample size of the control group to the treatment group). For attaining a specified power $1 - \beta$ and significance level α 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

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

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

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

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

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

respectively, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, $\mu_0 = d_0 \gamma_0$ is the average read count in the control group, d_0 is the average sequencing depth in the control group, and γ_0 is the control group gene expression level. In practice, for detecting a desired fold change ρ , we first must specify significance level α and power $1 - \beta$. Then, the ratio of the total number of reads between the two groups w and the average read count in the control group μ_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

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

where $p(x_0, x_1)$ is the p -value calculated by using (9) and $I(\cdot)$ denotes the indicator function. Required sample size n_{lr} then can be calculated by solving (15) with respect to n through a numerical approach, such as a gradient-search or bisection procedure.

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. Jung (2005) incorporated FDR controlling based on two-sample t-test under Gaussian distribution assumption. In this section, we borrowed their concept to incorporate the test statistics, such as Wald, Score, and likelihood ratio statistics, described in Section 2.2 for RNA-seq count data under the Poisson distribution assumption.

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 the expected proportion of false discoveries among rejected null hypotheses. Storey (2002) further proposed an improvement to FDR to achieve higher power, in the form

$$FDR = E\left(\frac{R_0}{R} \mid R > 0\right), \quad (16)$$

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 and defined it as (Storey, 2002; Storey and Tibshirani, 2001)

$$FDR = \frac{m_0\alpha}{m_0\alpha + E(R_1)}, \quad (17)$$

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

$$E(R_1) = \sum_{j \in M_1} \xi_j(\alpha), \quad (18)$$

where $\xi_j(\cdot)$ is the power function of the single α -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 (18) 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 (17) and (18) as

$$f = \frac{m_0\alpha}{m_0\alpha + r_1}, \quad (19)$$

and

$$r_1 = \sum_{j \in M_1} \xi(\rho_j, \mu_{0j}, \alpha, w, n). \quad (20)$$

By solving equation (19) 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 (20), 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, 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 ρ_j and average read count μ_{0j} prior to the RNA-seq experiment. Therefore, for getting a conservative sample size, we suggest using a common $\rho^* = \arg \min_{j \in M_1} \{|\rho_j - 1|\}$, minimum fold change, and $\mu_0^* = \min_{j \in M_1} \{\mu_{0j}\}$, minimum average read count, to estimate each ρ_j and μ_{0j} , respectively.

When we use ρ^* and μ_0^* in equation $g_1(n)$ to estimate each ρ_j and μ_{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, 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 sample size formulas (10)–(15) derived in Section 2.3 can be applied in the multiple-testing case, with the replacement of α and β with α^* and β^* . This reduces to the sample size calculation when we want to detect a fold change of ρ^* 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 of 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 μ_{0j} and ρ_j for each gene are known, use a numerical approach to solve the equation below with respect to n .

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

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

- b.** Otherwise,

- I.** specify a desired minimum fold change ρ^* and a minimum average read count μ_0^* ;
- II.** replace $\rho = \rho^*$, $\mu_0 = \mu_0^*$, $\varepsilon = r_1 f / (m_0(1 - f))$, 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 formulas under a finite sample and fixed alternative hypothesis, we conducted two-part numerical studies. In the first simulation, we focused on the single-gene testing problem and calculated the required sample size. In the second simulation, we focused on the multiple-gene testing problem which integrates the FDR issue into the sample size calculation procedure.

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

First, we focused on the single-gene testing problem. We assumed 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 log₂-fold change ($|\log_2(\rho)|$) are $0.32, 0.42, 0.58$ and 1.00). For each combination of design settings, we calculated the required sample size using our derived formulas, and then 5000 simulation samples of size n were generated from the same setting. The test statistics in (2)–(7) 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 $\alpha = 0.05$. 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 μ_0 increases. This result is as expected; for a fixed fold change, a small average read count provides less information, such that a larger sample size

is required to detect the difference. Moreover, for a fixed μ_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 each design settings, all empirical powers corresponding to n are close to nominal power 0.8. The results indicate the power of our methods is achieving the desired level. With regard to n , no method performed consistently 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_{lw} is smaller than others; for $w = 1$ and $\rho < 1$, n_{lp} is smaller than others; for $w = 1$ and $\rho > 1$, n_{ls} is smaller than others; for $w > 1$ and $\rho < 1$, n_{lp} is smaller than others; and for $w > 1$ and $\rho > 1$, n_w is smaller than others.

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

In the second simulation study, we focused on the multiple-gene testing problem and evaluated the performance of the sample size formulas while controlling FDR. We set the total number of genes for testing to $m = 4000$ and the number of statistically significant prognostic genes to $m_1 = 40$. We wanted to detect the expected number of true rejections $r_1 = 32$, which corresponds to a power of 80%. All parameters μ_j and ρ_j ($j = 1, \dots, 4000$) were assumed to be unknown. Thus, we used a minimum fold change ρ^* and a minimum average read count μ_0 to estimate each ρ_j and μ_j , $j = 1, \dots, 4000$. We varied $\mu^* = 1, 5$, 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). With these settings, $\alpha^* = 4.253 \times 10^{-4}$ and $\beta^* = 0.2$, which were calculated for controlling FDR at level 0.05 using the sample size calculation procedure described in Section 3. Then, we substituted α^* and β^* into the formulas (10)–(15) to calculate sample size under each method. In addition, for each design setting, we generated 5000 samples from independent Poisson distributions based on the calculated sample size n . The number of true rejections was counted using the q-value procedure proposed by Storey and Tibshirani (2003). The expected number of true rejections was estimated as the sample mean of the number of rejections of the 5000 simulation samples (r_1). In Table 2, we showed the calculated sample size with corresponding r_1 in parentheses under the case $w = 1$. The results for $w = 0.5$ or 2.0 are shown in the supplementary materials.

As seen in Table 2, all 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 the 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. Within a particular setting, however, each method tended to show consistently 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 the choice of sample size calculation to achieve the minimum required sample size.

5 Applications

5.1 Liver and kidney RNA-seq data set

Marioni et al. (2008) sequenced 7 liver and kidney RNA samples using Illumina's sequencing technology at two different cDNA concentrations (i.e. 3 pM of cDNA in five lanes and 1.5 pM of cDNA in two lanes). To eliminate the effects of cDNA concentration, they only used the samples with 3 pM of cDNA to identify differentially expressed genes between human liver and kidney RNA samples. Thus, we used the data set containing 5

human kidney samples and 5 human liver samples as pilot data in 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 and kidney samples, there were 17,306 genes left. We assumed that the top $m_1 = 175$ ($\approx 1\%$ of 17,306) genes, which have the largest absolute log2-fold change, are prognostic. From the pilot data, the minimum average read counts among the prognostic genes in the control group was estimated as $\mu_0^* = 5$. Suppose we want to identify 80% of the prognostic genes (i.e. $r_1 = 0.8 * 175 = 140$), while controlling the FDR at $f = 1\%$. Based on the pilot data, we set $m = 17,306$, $m_1 = 175$, $r_1 = 140$, and $f = 0.01$. In this case, we have

$$\alpha^* = \frac{140 \times 0.01}{17131 \times (1 - 0.01)} = 8.2549 \times 10^{-5}$$

and

$$\beta^* = 1 - \frac{140}{175} = 0.2.$$

Table 4 reports the sample size calculated from equations (10)–(15) under various desired minimum fold changes $\rho^* = 0.1, 0.25, 0.50, 0.75, 1.25, 1.50, 2.00, 2.50$ and 3.00 . From Table 4, we found that the original RNA-seq experiment described in Marioni et al. (2008) 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 3.00.

5.2 Transcript regulation data set

Blekhman et al. (2010) used RNA-seq to study transcript regulation in humans, chimpanzees, and rhesus macaques 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 13,267 genes in the data set after performing quality control analyses. Suppose that the top 133 ($\approx 1\%$ of 13,267) 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. From the pilot data, the minimum average read counts among the prognostic genes in the control group were estimated as $\mu_0^* = 1.67$ and the ratio of the total number of reads 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 $m = 13,267$, $m_1 = 133$, $m_0 = 13,134$, $r_1 = 107$, and $f = 0.1$. In this case, we have $\alpha^* = 9.0512 \times 10^{-4}$ and $\beta^* = 0.2$. Table 5 reports the sample size under various desired minimum fold changes while controlling the FDR at 10%. Suppose a new study is considered to detect differential gene expression at least two fold up- or down-regulated (i.e. $\rho^* = 2$ or 0.5). In such cases, we found that the sample size required is at most 62.

5.3 SAGE profiles of normal and tumor tissue

Zhang et al. (1997) examined human colorectal and pancreatic cancer tumor tissue. In this example, we used the data containing 2 colon tumor and 2 colon normal RNA samples as pilot data in designing a new study for comparing the profiles of normal and tumor tissue. There were 55,200 tags in the dataset. To keep tags that are expressed in reasonable counts, we filtered out the tags with no more than 5 counts. This reduces the dataset to 1193 tags. We assumed that the top $m_1 = 119$ ($\approx 1\%$ of 1193) tags, which have the largest absolute

log2-fold change, are prognostic. From the filtered dataset, the minimum average read counts among the prognostic tags in the normal tissue group were estimated as $\mu_0^* = 5.0$ and the ratio of the total number of reads between the two groups was estimated as $w = 0.85$. Suppose we want to identify 85% of the prognostic tags (i.e. $r_1 = 0.85 \times 119 = 101$), while controlling the FDR at 5%. Based on the pilot data, we set $m = 1193$, $m_1 = 119$, $m_0 = 1074$, $r_1 = 101$, and $f = 0.05$. In this case, we have $\alpha^* = 0.005$ and $\beta^* = 0.15$. Table 6 reports the sample size under various desired minimum fold changes while controlling the FDR at 5%. For instance, suppose that a new study allows a data collection plan to be implemented with a sample size less than 50 due to cost considerations. From Table 6, we found that a detectable minimum up- or down-regulation fold change is 1.5 (i.e. $\rho^* = 1.5$ or 0.67) in such situation.

6 Conclusion

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

In this paper, because RNA-seq data are count data, we assume the read counts follow a Poisson distribution naturally. A number of testing methods based on a Poisson distribution have been developed. For instance, Marioni et al. (2008) proposed a Poisson log-linear model and utilized the likelihood ratio test; Wang et al. (2010b) assumed that log ratios of counts have a normal distribution and utilized z-score; Li et al. (2011) proposed a Poisson log-linear model and utilized the score test. 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 (Robinson and Smyth, 2008). When the read counts exhibit overdispersion (i.e. variation greater than mean), the sample size calculation based on the Poisson model will be underestimated due to underestimation of the variance. In this case, the study based on the corresponding sample size will be underpowered. To handle data with over-dispersion, Li et al. (2011) suggested using power transformation to make the data follow a Poisson distribution more closely. For data with overdispersion, therefore, our sample size calculation methods based on the transformed data could allow for a wider applicability. On the other hand, to model data with overdispersion directly, several approaches for testing differential gene expression in RNAseq data have been proposed (e.g., negative binomial model, Anders and Huber, 2010; generalized Poisson model, Srivastava and Chen, 2010). The extension of our sample size calculation procedure to those approaches will be left for future work. In addition, the matched pairs design is a common case in clinical research. However, the proposed method cannot account for the correlation between paired samples.

Thus, the calculation of sample size for a matched pairs design would be an interesting research topic for us in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Lynne Berry and Margot Bjoring for their editorial support. Special thanks also go to the editor and three anonymous reviewers for their valuable comments and suggestions.

Biographies

Chung-I Li received the PhD degree in Statistics from National Cheng Kung University in Taiwan. He is currently a postdoctoral research fellow of Center for Quantitative Sciences at Vanderbilt University. His research interests include statistical quality control and RNA-seq data analysis.

Pei-Fang Su received the PhD degree in Statistics from National Cheng Kung University in Taiwan. She is a postdoctoral research fellow of Center for Quantitative Sciences at Vanderbilt University. Her research interests include statistical bioinformatic analysis of high-dimensional omics data and statistical methods for survival analysis and linear models.

Yan Guo is a research Assistant Professor of Department of Cancer Biology at Vanderbilt University. He received his PhD in Computer Science from University of South Carolina.

Yu Shyr received his PhD in Biostatistics from University of Michigan (Ann Arbor) in 1994 and subsequently joined the faculty at Vanderbilt University School of Medicine. He serves as the director of Center for Quantitative Sciences at Vanderbilt University, Direct of Technologies for Advanced Genomics Analysis Research Design, associate director for Quantitative Sciences Integration, and professor in Department of Biostatistics at Vanderbilt University. His current research interests lie in developing and analysing predictive models of the statistical relationships between multiple-variable protein and RNA expression data and clinical endpoints using both supervised and unsupervised classification and pattern recognition approaches.

References

- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010; Vol. 11(No. 10):R106. [PubMed: 20979621]
- Auer PL, Doerge RW. Statistical design and analysis of rna sequencing data. *Genetics.* 2010; Vol. 185(No. 2):405–416. [PubMed: 20439781]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.* 1995; Vol. 57(No. 1):289–300.
- Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* 2010; Vol. 20(No. 2):180–189. [PubMed: 20009012]
- 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; Vol. 5(No. 7):613–619. [PubMed: 18516046]
- Fang Z, Cui X. Design and validation issues in rna-seq experiments. *Brief. Bioinform.* 2011; Vol. 12(No. 3):280–287. [PubMed: 21498551]

- Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou Y-C, 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; Vol. 7(No. 10):843–847. [PubMed: 20835245]
- Gu K, Ng HKT, Tang ML, Schucany WR. Testing the ratio of two poisson rates. *Biom. J.* 2008; Vol. 50(No. 2):283–298. [PubMed: 18311854]
- Hashimoto, S-i; 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; Vol. 4(No. 1):e4108. [PubMed: 19119315]
- 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; Vol. 3:140–148.
- Hu J, Zou F, Wright FA. Practical fdr-based sample size calculations in microarray experiment. *Bioinformatics*. 2005; Vol. 21:3264–3272. [PubMed: 15932903]
- Huffman MD. An improved approximate two-sample poisson test. *J. R. Stat. Soc. Ser. B.* 1984; Vol. 33:224–226.
- Jiang H, Wong WH. Statistical inferences for isoform expression in rnaseq. *Bioinformatics*. 2009; Vol. 25(No. 8):1026–1032. [PubMed: 19244387]
- Jung S-H. Sample size for fdr-control in microarray data analysis. *Bioinformatics*. 2005; Vol. 21(No. 14):3097–3104. [PubMed: 15845654]
- Krishnamoorhy K, Thomson J. A more powerful test for comparing two poisson means. *J. Stat. Plan Infer.* 2004; Vol. 119:23–35.
- Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. Rnaseq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2010; Vol. 26(No. 4):493–500. [PubMed: 20022975]
- Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*. 2011; Vol. 14:1–16.
- Liu P, Hwang JTG. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*. 2007; Vol. 23(No. 6):739–746. [PubMed: 17237060]
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. Rnaseq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008; Vol. 18(No. 9):1509–1517. [PubMed: 18550803]
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat. Methods*. 2008; Vol. 5(No. 7):621–628. [PubMed: 18516045]
- Ng HKT, Tang M-L. Testing the equality of two poisson means using the rate ratio. *Stat. Med.* 2005; Vol. 24(No. 6):955–965. [PubMed: 15532090]
- Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005; Vol. 21:3017–3024. [PubMed: 15840707]
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*. 2010; Vol. 464(No. 7289):768–772. [PubMed: 20220758]
- Pounds S, Cheng C. Sample size determination for the false discovery rate. *Bioinformatics*. 2005; Vol. 21(No. 23):4263–4271. [PubMed: 16204346]
- 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 A-L, 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; Vol. 7(No. 11):909–912. [PubMed: 20935650]
- Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*. 2008; Vol. 9(No. 2):321–332. [PubMed: 17728317]

- Srivastava S, Chen L. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Res.* 2010; Vol. 38(No. 17):e170. [PubMed: 20671027]
- Storey JD. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B.* 2002; Vol. 64(No. 3): 479–498.
- Storey, JD.; Tibshirani, R. Technical report. Stanford University, CA: Department of Statistics; 2001. Estimating false discovery rates under dependence, with applications to dna microarrays.
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 2003; Vol. 100(No. 16):9440–9445. [PubMed: 12883005]
- 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; Vol. 28(No. 5):511–515. [PubMed: 20436464]
- Wang L, Feng Z, Wang X, Wang X, Zhang X. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics.* 2010a; Vol. 26(No. 1):136–138. [PubMed: 19855105]
- 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.* 2010b; Vol. 5(No. 1):e8529. [PubMed: 20072613]
- Wu Z, Wang X, Zhang X. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics.* 2011; Vol. 27(No. 4):502–508. [PubMed: 21169371]
- Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. Gene expression profiles in normal and cancer cells. *Science.* 1997; Vol. 276(No. 5316):1268–1272. [PubMed: 9157888]

Table 1

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

w	ρ ($ \log_2(\rho) $)	μ_0	n_w	n_s	n_{hw}	n_{ls}	n_{lp}	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	1	173 (.81)	162 (.80)	159 (.80)	167 (.80)	156 (.79)	164 (.80)

w	ρ ($ \log_2(\rho) $)	μ_0	n_w	n_s	n_{lw}	n_{ls}	n_{lp}	n_{lr}
	(0.42)	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	1	55 (.80)	61 (.81)	64 (.82)	57 (.81)	64 (.83)	64 (.83)
	(0.58)	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	40 (.83)	34 (.81)	33 (.80)	36 (.80)	31 (.78)	35 (.81)
	(1.00)	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 r_1) for $r_1 = 32$ at FDR = 0.05 when $w = 1$, $m = 4000$, $m_1 = 40$

$ \log_2(\rho) $	μ_0	n_w	n_s	n_{lw}	n_{ls}	n_{lp}	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_{lp}	n_{ls}
$w > 1$	n_{lp}	n_w

Table 4

Sample size calculation under various desired minimum fold changes (ρ^*) for $r_1 = 140$, at FDR = 1% when $m = 17,306$ and $m_1 = 175$

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

Table 5

Sample size calculation under various desired minimum fold changes (ρ^*) for $r_1 = 107$, at FDR = 1% when $m = 13,267$ and $m_1 = 133$

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

Table 6

Sample size calculation under various desired minimum fold changes (ρ^*) for $r_1 = 101$, at $FDR = 5\%$ when $m = 1193$ and $m_1 = 119$

ρ^*	n_w	n_s	n_{lw}	n_{ls}	n_{lp}	n_{lr}
0.10	5	5	8	4	3	6
0.25	7	8	9	6	6	7
0.50	19	20	21	19	18	19
0.67	48	49	50	48	46	48
1.25	117	116	116	116	119	120
1.50	33	32	33	32	34	33
2.00	10	10	10	10	11	10
2.50	6	5	6	5	6	5
3.00	4	4	4	4	3	4