RSPS: an R library for Empirical Determination of Statistical Power for RNA-Seq Studies

Milan Bimali[1], Joseph Usset [1], Brooke Fridley[1]

[1] Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS 66160, USA

Abstract

Summary: Next generation sequencing technology is a powerful technology that enables researchers to discover, profile, and quantify transcripts across the entire transcriptome. Benefits of RNA-Seq over microarray technologies include: the ability to assess alternative splicing; detection of gene fusions; improved dynamic range; and the ability to use on non-model organisms. A fundamental question that arises in the design of many RNA-Seq studies is the required sample size to achieve a desired statistical power to determine differentially expressed genes. Here, we present an R package *RSPS* (RNA-Seq Power Simulation) that uses an efficient simulation algorithm to empirically determine statistical power or the necessary sample size for an RNA-Seq study.
Availability: We have uploaded an R package *RSPS* in R CRAN which computes sample size estimates based on empirical simulation estimates.
Contact: bfridley@kumc.edu

Key Words: RNA-Seq, Sample Size Estimation, Simulation

1. Introduction

Next-generation sequencing of transcriptomes (i.e., RNA-Seq) quantifies the transcriptome at a maximal resolution and dynamic range and without any prior assumption or biological knowledge of the organism (1). Unlike the microarray technology which can only quantify relative gene expression levels, RNA-Seq can quantify absolute gene expression levels in addition to determining gene fusions, alternative splicing and novel isoforms, and other expressed genetic variants. Many of the analyses approaches for microarray based gene expression studies have been carried forward into the analysis of RNA-Seq data, including sample sizes calculations based on the traditional two-sample t-test (i.e. normal distribution). However, a key challenge is that RNA-Seq produces count data that do not follow a Gaussian distribution, but rather an over-dispersed Poisson or Negative Binomial distribution, thereby making the use of t-test based sample size calculation inappropriate as it is assuming an incorrect error distribution (2, 3).

To date, several methods to calculate sample sizes for RNA-Seq studies have been proposed based on large sample asymptotics of the Poisson or Negative Binomial based test statistics. For example, Li et al. derive sample size calculations based on asymptotic statistics under an assumed Poisson model for RNA-Seq data (4). However, a Poisson model assumes an equal mean and variance of RNA-seq read counts. Often read counts exhibit over-dispersion (i.e. variance greater than the mean) and in this case necessary sample sizes calculated based on a Poisson distribution will be underestimated. To account for over-dispersion several power calculation methods have been proposed based on a Negative Binomial model (5, 6). However, these methods rely on generalized linear models for the Negative Binomial model for which for which analytical solutions do not exist.

A primary limitation of the proposed approaches are that they are based on normality approximations to test statistics based on generalized linear models with Poisson or Negative Binomial distribution errors. Hence, the performance of these methods for small studies is suspicious. This is particularly the case in many basic science experiments with limited sample sizes (i.e., mouse studies, xenograft studies). Consequently, there exists a need for sample size calculation software that accurately estimates power for RNA-seq studies with small sample sizes.

The primary aim of this R package *RSPS* (RNA-Seq Power Simulation) is to provide a simulation-based sample size or power determination for RNA-seq studies to determine differential expressed genes between two groups or conditions. *RSPS* has an advantage over the proposed methods in that it does not rely on normality approximations to test statistics based on the Poisson or Negative Binomial distributions based on either asymptotic arguments or data transformations. RSPS allows the user to specify either an underlying Poisson or Negative Binomial model for the sequencing data; can estimate necessary sample size for a desired power (given effect size) or vice versa; and provides power curves and tables that could be used within grant applications or protocols.

The R package produced by the project is freely available from CRAN.

2. Implementation

Estimating the necessary sample size for an experiment generally requires four factors: the level of Type I error (α), power of the test $(1-\beta)$ or equivalently the Type II error rate (β), an estimate of variability in the outcome of interest, and an estimate of clinically significant differences in mean log-fold change one wishes to be able to detect. In theory if any two of three parameters are known/fixed (assuming the Type I error is pre-specified), the third value can be computed. However in practice, the estimate of effect size or fold change is usually assumed and varied over a range of possible values; with either the power being estimated for a given sample size or the sample size being estimated to achieve the desired power is computed.

In RSPS, the user inputs a nominal Type-I error rate, a vector of mean-fold changes (effect sizes of interest), and the underlying data generative distribution (Poisson or Negative Binomial). If the Negative Binomial distribution is specified the user also inputs an estimate of the over-dispersion parameter that reflects the mean-variance relationship in the data. Let N_{gi} be the number of RNA-seq counts for group g and subject i. Then for the negative binomial parametrization, the count data is generated from:

$$\log \lambda_g = \alpha + \beta_g ; \quad Var(N_{gi}) = \lambda_{gi} + \phi \lambda_{gi}^2 ,$$

where λ_g is the mean counts for group g, β_g is the fold change, and ϕ is the over-dispersion parameter. For the Poisson distribution $\phi = 0$. Depending on the goals of the empirical simulation study, the user inputs either a tentative sample size or a desired power.

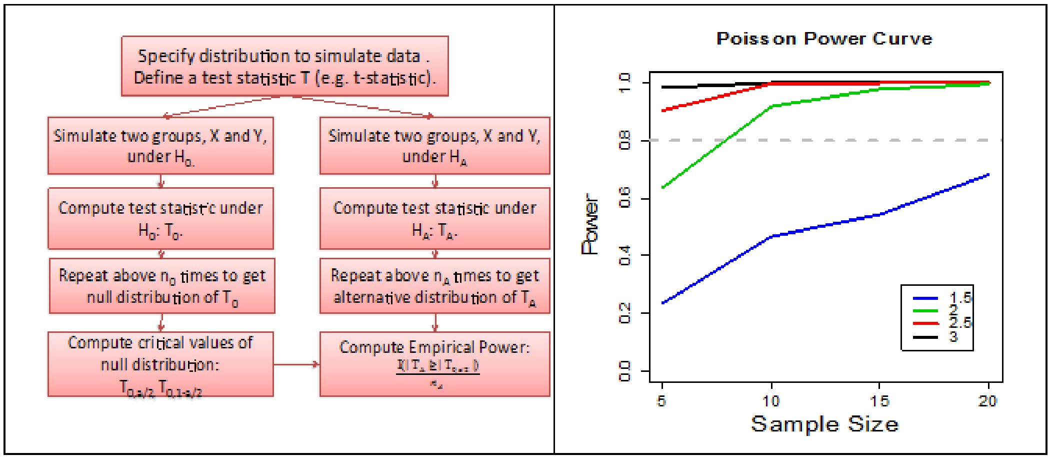

Figure 1: a. The empirical simulation algorithm. **b.** Power curve output by RSPS.

The algorithm for simulation based power estimation proposed in RSPS is shown in **Figure 1a**. The algorithms works as follows. First, RSPS simulates data RNA-Seq data from two groups according to the specified underlying data distribution, under the null hypothesis that the log-fold mean change between the groups is 0. For each data set simulated under the null hypothesis a two-sample t-statistic is calculated. This process is repeated n_0 (5000 pre-set) to construct a null-distribution of the underlying t-statistic. From this simulated null distribution the critical values for hypothesis testing are obtained. Next, the data are generated under the alternative hypothesis that the two groups have different mean log-fold counts. The difference in the magnitude of the mean log-fold counts between the two groups under the alternative hypothesis is pre-specified by the user. For each simulated data set under the alternative hypothesis a two-sample t-test of the log-mean shift between the two groups is calculated. This process is repeated n_A times. The proportion of the test-statistics generated under the alternative hypothesis more extreme than the empirical null distribution critical values represent an estimate of the power.

RSPS estimates the sample size given a desired power similarly to the algorithm above using a grid search. Specifically, power is calculated across a variety of sample sizes; and then output the minimum of these sample sizes required to achieve the desired power. Example output (power curves) are presented in **Figure 1b**. While the grid search requires additional computation RSPS it still relatively fast and can calculate 1000 simulations across 5 simulation values in under 10 seconds.

Conclusions

Sample size and power estimation is a fundamental step in the design of any experiment to ensure that you have the best possible chance of determining the truth. The digital count nature of RNA-Seq data, make it unsuited for standard testing methods that assume the data is normally distributed, particular when dealing with relatively small studies. Previous approaches to power and sample size estimates based on the distributions rely on data transformations or asymptotic arguments to justify normality assumptions on the test statistics measuring differential expression. However, in our experience, many differential expression analyses the have small sample sizes and therefore asymptotic approximations are poor. The main advantage of RSPS is that sample size and power calculation do not rely on asymptotic normality approximations, and therefore should be more appropriate for studies with small sample size.

RSPS currently uses a t-statistic within the simulation algorithm to generate distribution between two groups, but is an ongoing project and will be generalized to generate the null distribution with other test statistics, and also extended to handle pair designs, continuous predictors. A limitation of our method is that in practice the p-value for the t-statistic will be based on asymptotic approximations of normality from the data sample. Therefore, RSPS is best suited to correspond to analyses where differential expression is assessed by re-sampling the data under the null and alternative hypotheses, and comparing the distribution of the estimated t-statistics from these re-sampling distributions.

References

1. Marguerat S, Bähler J. RNA-seq: from technology to biology. Cellular and molecular life sciences. 2010;67:569-79.
2. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010;11:R106.
3. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology. 2010;11:R25.
4. Li C-I, Su P-F, Guo Y, Shyr Y. Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution. International journal of computational biology and drug design. 2013;6:358-75.
5. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. rna. 2014;20:1684-96.
6. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. Calculating sample size estimates for RNA sequencing data. Journal of Computational Biology. 2013;20:970-8.