Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Network and Pathway Analysis of Cancer Susceptibility (B)

RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment

Yan Guo[1,*], Shilin Zhao[1,*], Chung-I Li[2], Quanhu Sheng[1] and Yu Shyr[1]

[1]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN, USA. [2]Department of Statistics, National Cheng Kung University, Tainan, Taiwan. *Equal contribution.

ABSTRACT: Sample size and power determination is the first step in the experimental design of a successful study. Sample size and power calculation is required for applications for National Institutes of Health (NIH) funding. Sample size and power calculation is well established for traditional biological studies such as mouse model, genome wide association study (GWAS), and microarray studies. Recent developments in high-throughput sequencing technology have allowed RNAseq to replace microarray as the technology of choice for high-throughput gene expression profiling. However, the sample size and power analysis of RNAseq technology is an underdeveloped area. Here, we present RNAseqPS, an advanced online RNAseq power and sample size calculation tool based on the Poisson and negative binomial distributions. RNAseqPS was built using the Shiny package in R. It provides an interactive graphical user interface that allows the users to easily conduct sample size and power analysis for RNAseq experimental design. RNAseqPS can be accessed directly at http://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/.

KEYWORDS: RNAseq, power analysis, sample size calculation, experiment design

SUPPLEMENT: Network and Pathway Analysis of Cancer Susceptibility (B)

CITATION: Guo et al. RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. *Cancer Informatics* 2014:13(S6) 1–5 doi: 10.4137/CIN.S17688.

RECEIVED: June 11, 2014. **RESUBMITTED:** July 18, 2014. **ACCEPTED FOR PUBLICATION:** July 22, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: This study was supported by grant CCSG (P30 CA068485). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: yan.guo@vanderbilt.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

Gene expression profiling has been an important component of biomedical research. Gene expression is the appearance of a characteristic or effect in the phenotype that can be attributed to a certain gene. For over a decade, microarray technology was the dominating technology for high-throughput gene expression profiling until the introduction of RNAseq technology. RNAseq technology is a form of next-generation sequencing (NGS) technology, sometimes referred to as high-throughput sequencing technology. Since its introduction, NGS technology has revolutionized genetic and biomedical research. RNAseq is the use of NGS technology to sequence cDNA (reversed transcribed from RNA) in order to obtain information about RNA. Compared to microarray technology, RNAseq technology offers several obvious advantages. First, RNAseq allows the detection of all isoforms of a gene, even novel ones. Microarray, on the other hand, relies purely on previous knowledge regarding genes to design probes for detection, thus it cannot be used for novel detection. Second, the resolution of microarray usually stays at the gene and exon level, but the resolution of RNAseq can reach the level of a single nucleotide, which allows detection of single nucleotide variance and structural variants, such as small insertions, deletions, alternative splicing, and gene fusion. Although the cost of RNAseq has become comparable to microarray, researchers have unanimously agreed that RNAseq has replaced microarray as the go to technology of high-throughput gene expression profiling.[1–4]

Although RNAseq technology has introduced exciting opportunities to biomedical researchers, it has also created several challenges for analysts. For example, the data storage cost is significantly higher compared to microarray. And the lack of consensus on the best statistical method for detecting differentially expressed genes[5] is really noticeable. Another easily overlooked area is sample size and power calculation for RNAseq-based experiment designs. Experiment design is a pivotal step and the first step of any successful study, and one of the most important questions to address during experiment design is what sample size is needed to achieve the desired statistical power within the financial budget. Sample size and power analyses are required for most National Institutes of Health (NIH) funding applications. For microarray-based studies, sample size and power calculation has been well established.[6–8] The sample size and power for microarray-based studies are relatively easy to compute, because gene expression of microarray data follows a normal distribution. However, RNAseq data are count based, which is usually modeled as a Poisson[9,10] or negative binomial distribution.[11,12] Several studies have attempted to produce methods to estimate sample size and power based on Poisson and negative binomial distributions for RNAseq-based experiment design. For example, Fang and Cui[13] derived a sample size calculation formula based on the Wald test for a single-gene differential expression test. Busby et al.[14] introduced Scotty, a web tool for computed sample size using a *t*-test for RNAseq data based on a Poisson-lognormal distribution. Hart et al.[15] introduced RNAseqPower, a method for sample size calculation using the score test based on a negative binomial distribution. R code and the excel worksheet for this method were provided by the authors. In reality, RNAseq experiment is capable of detecting thousands of genes, and those genes are tested for the significance of differential expression simultaneously. In such cases, the correction of error rates for multiple comparisons is required. However, both Scotty and RNAseqPower fail to discuss how to calculate sample size under multiple comparison testing. To address this issue, Li et al.[16] derived a sample size calculation formula based on the Poisson distribution. Later, Li et al.[17] proposed a sample size calculation method based on the exact test.[18] Of all the developed methods, only Scotty has a web-based graphical user interface. The other proposed methods require running R code. A convenient, user-friendly RNAseq sample size and power calculation tool is highly desirable.

Here, we present RNAseqPS, a web-based power and sample size calculation tool, based on both the Poisson and negative binomial distributions. RNAseqPS is based on the methods[16,17] developed by Li et al. RNAseqPS distinguishes itself from other RNAseq sample size and power calculation methods in several aspects. First, the graphical user interface of RNAseqPS is highly interactive and intuitive. It offers several different sample size and power calculation methods within in one interface. Also, it does not require the skill of running R code, a person without any previous experience in R can operate it with ease. There are other features of RNAseqPS which are discussed in detail in the Method section.

Method

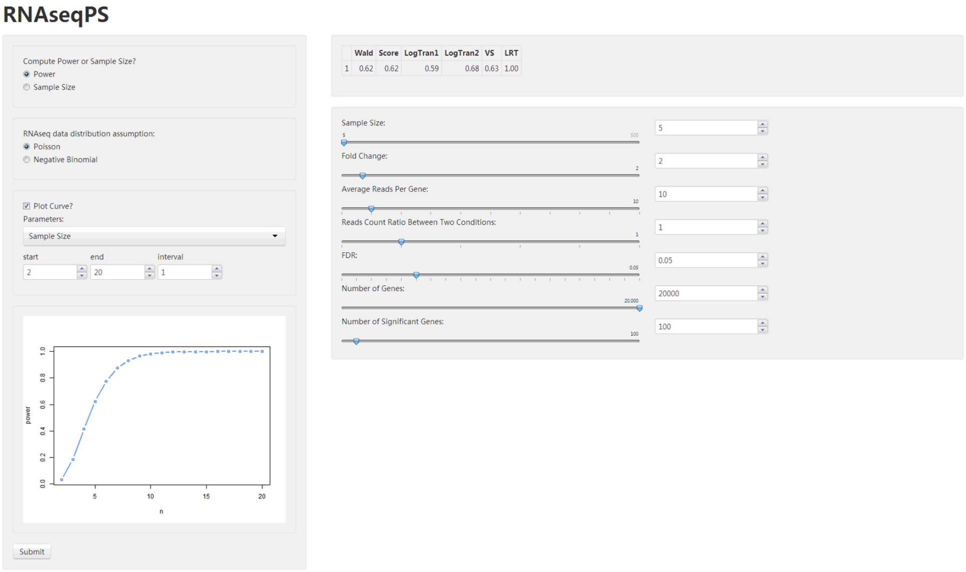
For the Poisson distribution, we considered six different methods: (1) Wald test; (2) score test; (3) likelihood ratio test; (4) log transformation of Wald statistic; (5) log transformation of score statistic; and (6) transformation of Poisson for calculating sample size under both single- and multiple-gene comparison scenarios. The detail of the implementation of these tests can be found.[16] Under the Poisson distribution, we assume that mean equals variance. When biological replicates are used, RNAseq data could exhibit variation significantly greater than the mean (over-dispersion). In such scenario, the Poisson distribution cannot model the data properly with over-dispersion. The negative binomial distribution can be used as a natural extension of the Poisson model. For the negative binomial distribution, we provide a method to calculate power for the exact test proposed.[18] based on a given *P* value under the single-gene comparison scenario. The detail of the implementation of this test can be found.[17] To address the multiple comparison issue, we further incorporated false discovery rate (FDR)[19] controlling into our methods described.[16,17] RNAseqPS implements sample size and power calculation from both Poisson and negative binomial distributions based on the methods proposed.[16,17]

Results

RNAseqPS is written using the Shiny package in R. Shiny is a powerful web framework for building interactive web applications using R. It combines the computation power of R with the interactivity of modern web. Applications built using Shiny can automatically react with the input and output parameters of an application. Thus, when new parameters are input into RNAseqPS, sample size or power is automatically recomputed and displayed on the screen. RNAseqPS has a very interactive and dynamic graphical user interface (Fig. 1). Users without any experiences in statistics and programing languages can navigate through it without trouble.

RNAseqPS also has the ability to plot the relationship of any parameter with power. For example, a user can specify sample size as the X-axis and selects a reasonable range for sample size. RNAseqPS can plot relationship of the sample size with power while other parameters stay unchanged. Such plots help users visualize the intrinsic relationship between the parameters and power and can be easily exported as high-resolution figures that can be used for NIH grant funding applications. Several examples of such plots can be seen in Figure 2.

Each computation of sample size or power takes up to 30 seconds to 1 minute of time. To reduce the user wait time, we have implemented a database lookup function with RNAseqPS. The database lookup function works as follows.



RNAseqPS

Compute Power or Sample Size?
● Power
○ Sample Size

RNAseq data distribution assumption:
● Poisson
○ Negative Binomial

☑ Plot Curve?
Parameters: Sample Size
start 2
end 20
interval 1

Submit

	Wald	Score	LogTran1	LogTran2	VS	LRT
1	0.62	0.62	0.59	0.68	0.63	1.00

Sample Size: 5
Fold Change: 2
Average Reads Per Gene: 10
Reads Count Ratio Between Two Conditions: 1
FDR: 0.05
Number of Genes: 20000
Number of Significant Genes: 100

Figure 1. Example of the graphical interface of RNAseqPS.

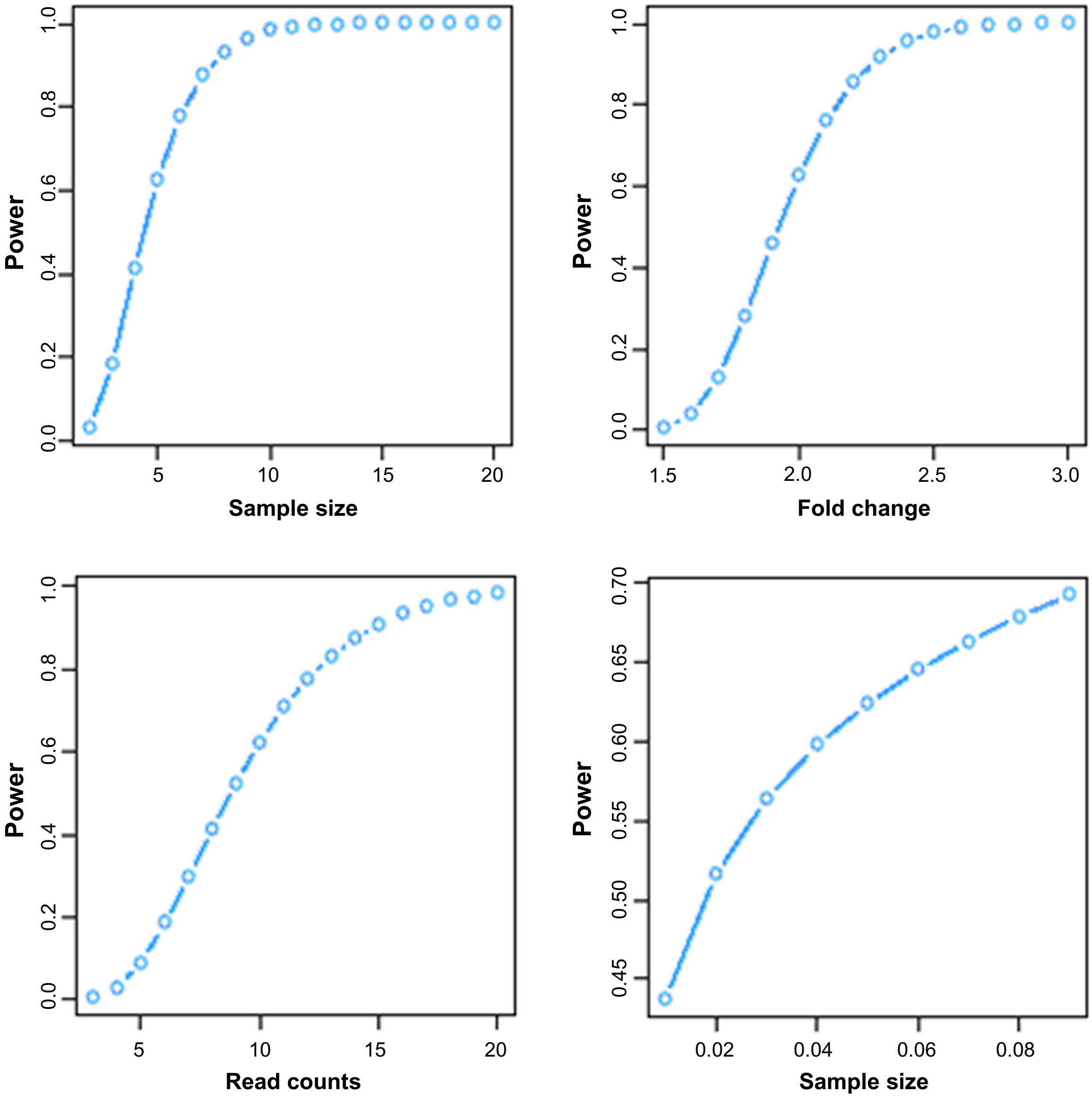


Figure 2. Examples of the power curves produced by RNAseqPS.

To estimate the power of the RNAseq experiment under the Poisson distribution, the following eight parameters are required as input: (1) sample size; (2) expected fold change; (3) average reads per gene; (4) read count ratio between two conditions; (5) FDR threshold; (6) total number of genes; (7) expected number of significant genes and (8) dispersion (only required when using negative binomial distribution). The details of these parameters are explained in Table 1. For each of these parameters, there are certain likely ranges. For example, for expected fold change parameter, the range is usually between absolute value 1 to 10 because the absolute value of the fold change cannot be less than 1 and fold change 10 is a reasonable assumption of the high ceiling. To produce the database, we pre-computed the power and sample size using the parameter values within these most likely ranges for all parameters and stored them in a My SQL database. For each round of input parameters, RNAseqPS first queries the database to see if samples size or power for such set of parameters has already been entered into the database. If the input parameters match the existing parameters in the database, the power or sample size is automatically returned, otherwise power or sample size will be computed on the fly. RNAseqPS allows two ways to input parameters: slide bar and text box. When using slide bar, the input is guaranteed to match an input in the database. Using the text box input, user can input any values as desired. In such cases, the sample size and power will be computed on the fly. By using the look up function, we have instant feedback for most of the common RNAseq experiment designs.

Discussion

The dominance of RNAseq over microarray in gene expression research has been documented by multiple studies.[1–4] Even though the price of RNAseq has matched that of microarray, large studies can still be costly. Insufficient sample size may lead to underpowered studies and produce unreliable results. Sample size estimation is a critical issue for any biological studies, including RNAseq-based studies. Better management of the tradeoff between cost and sample size is key to a successful study. Sample size and power analysis for RNAseq is an underdeveloped area when compared with microarray. In comparison, sample size and power is easier to calculate for microarray, because microarray follows a normal distribution, while RNAseq data are count based and follow a Poisson or negative binomial distribution. We have discussed several sample size and power calculation methods developed previously. These methods adequately address the sample size and power calculation needs for an RNAseq experiment design but still have two issues. First, some of them do not adjust for the multiple comparisons problem. Second, a majority of them lack an intuitive graphical user interface.

We address these two issues with RNAseqPS. RNAseqPS addresses the multiple comparisons problem by implementing the methods introduced by Li et al., which incorporate FDR controls. Also, RNAseqPS implements methods based on both the Poisson and negative binomial distributions. Overall, there are seven methods implemented in RNAseqPS. Users can select appropriate methods based on the experiment design. Finally, RNAseqPS offers a highly interactive and user-friendly graphical user interface. This graphical user interface helps users input the parameters with ease and greatly enhances users' abilities to understand the sample size and power analysis with the help of visualization. Overall, RNAseqPS addresses major issues currently lacking in sample size and power analysis for RNAseq experiment design. It greatly benefits researchers who plan to use RNAseq as the primary technology for their study.

Acknowledgement

The authors thank Margot Bjoring for her editorial support.

Table 1. RNAseqPS input parameters.

PARAMETERS	LOWER BOUND	UPPER BOUND	INTERVAL	NOTE
Sample size	1	500	1	Required when computing power
Desired power	0.8	0.95	0.05	Required when computing sample size. The minimum power should be no less than 80%
Expected fold change	1.4	10	0.2	The expected fold change between differentially expressed genes. This value is based on prior experience. If no previous data is available, a best guess is given by RNAseqPS
Average reads per gene	1	100	10	This can be computed as R/G, where R is the total number of reads sequenced and G is the total number of genes detected
Total number genes	100	20000	100	This is usually dependent on the gene transfer format (GTF) file used. A GTF file contains the annotation information regarding genes, and it is required for RNAseq analysis
Expected number of differentially expressed genes	5	2000	50	This is the number of genes you expect to see between the two conditions. It is also based on prior knowledge. When prior knowledge is unavailable, a best guess is provided by RNAseqPS
Dispersion	0.1	2	0.1	This parameter is used in the negative binomial model

Author Contributions

YG wrote the manuscript. YG and SZ programmed the user interface. QS helped with the software programming. CL and YS designed the algorithm for RNAseq power analysis and sample size calculation. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
2. Shendure J. The beginning of the end for microarrays? *Nat Methods*. 2008;5(7):585–7.
3. Guo Y, Li CI, Ye F, Shyr Y. Evaluation of read count based RNAseq analysis methods. *BMC Genomics*. 2013;14(suppl 8):S2.
4. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One*. 2013;8(8):e71462.
5. Guo Y, Zhao S, Ye F, Sheng Q, Shyr Y. MultiRankSeq: multiperspective approach for RNAseq differential expression analysis and quality control. *Biomed Res Int*. 2014;2014:8.
6. Lee ML, Whitmore GA. Power and sample size for DNA microarray studies. *Stat Med*. 2002;21(23):3543–70.
7. Lin WJ, Hsueh HM, Chen JJ. Power and sample size estimation in microarray studies. *BMC Bioinformatics*. 2010;11:48.
8. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*. 2005;6(1):27–38.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
10. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 2010;26(1):136–8.
11. Di Y, Schafer D, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol*. 2011;10(1):1–28.
12. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.
13. Fang Z, Cui X. Design and validation issues in RNA-seq experiments. *Brief Bioinform*. 2011;12(3):280–7.
14. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*. 2013;29(5):656–7.
15. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. *J Comput Biol*. 2013;20(12):970–8.
16. Li CI, Su PF, Guo Y, Shyr Y. Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution. *Int J Comput Biol Drug Des*. 2013;6(4):358–75.
17. Li CI, Su PF, Shyr Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics*. 2013;14:357.
18. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23(21):2881–7.
19. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol*. 2002;64:479–98.