Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Research Article Open Access

Sample Size Calculation of RNA-sequencing Experiment-A Simulation-Based Approach of TCGA Data

Derek Shyr[1] and Chung-I Li[2]*

[1]*Washington University, St. Louis, MO 63130, USA*
[2]*Department of Applied Mathematics, National Chiayi University, Chiayi, Taiwan 60004, ROC*

Abstract

Power and sample size calculation is an essential component of experimental design in biomedical research. For RNA-sequencing experiments, sample size calculations have been proposed based on mathematical models such as Poisson and negative binomial; however, RNA-seq data has exhibited variations, i.e. over-dispersion, that has caused past calculation methods to be under- or over-power. Because of this issue and the field's lack of a simulation-based sample size calculation method for assessing differential expression analysis of RNA-seq data, we developed this method and applied it to three cancer sites from the Tumor Cancer Genome Atlas. Our results showed that each cancer site had its own unique dispersion distribution, which influenced the power and sample size calculation.

Keywords: Next generation sequencing; RNA-sequencing; Tumor cancer genome atlas; Cancer; sample size; Power; Simulation

Abbreviations: NGS: Next-generation sequencing; RNA-seq: RNA-sequencing; TCGA: Tumor Cancer Genome Atlas

Introduction

NGS has given scientists the ability to characterize and view the genome at a high resolution. With the development of several NGS applications, including whole-genome, whole-exome, chromatin immunoprecipitation, and others, scientists have uncovered various mutations and variations in the genome, such as point mutations, small indels, copy number variations, gene fusions, alternative splicing, etc. As NGS continues to produce an enormous volume of data at a fast rate and economic price, scientists are currently applying these data to experiments to not only gain a better understanding of certain biomarkers and genes, but also discover possible target therapies for diseases like cancer [1-4].

Scientists are publishing their results among the top journals of the world and providing new ideas for drug development. While the idea of translating research to bedside therapies is ideal, clinical success has been particularly low for cancer. Compared with other therapeutic areas, cancer clinical trials have the highest failure rate. These failures have been attributed to the quality of published preclinical data given that drug development relies heavily on these findings. Among fifty-three cancer papers that were published in high-impacting journals and known as "landmark" studies, only six of them were reproducible [5,6]. Given these results, there's a strong need to raise the quality of preclinical studies by having more rigor in experimental design. Among the six reproducible results, the studies paid attention to bias, controls, randomization, and other important factors that can make an impact on the reliability of the results [5,6].

A lack of understanding the statistical concepts of type I error and type II error has also contributed to the number of irreproducible results. While scientists emphasize the importance of having a low type I error probability or the false positives probability, some do not realize that type II error or the false negatives also play a significant factor in determining the outcome of an experiment. The probability of true positives, also known as the power, is equivalent to (1−type II error probability); thus, in order to ensure that most of the experiments' significant findings are actually correct, the type II error probability must be low [6-8]. Power of the test depends on the number of subjects assigned in an experiment. As one increases the number of subjects, the amount of power increases [7].

In NGS research, RNA-seq has proven to be useful to many researchers who have generated multiple research questions from discovering and profiling RNA transcripts with novel transcripts, alternative splicing, and other variations [9]. Because thousands of genes are examined in a RNA-seq experiment, differential expression among those genes is tested simultaneously, requiring the correction of error rates for multiple comparisons. For the high-dimensional multiple testing problem, several such corrected measures have been proposed, such as family-wise error rate (FWER) and false discovery rate (FDR). In high-dimensional multiple testing circumstances, controlling FDR is preferable because the Bonferroni correction for FWER is often too conservative. For the multiple testing problem, the FDR is defined as

$$\mathrm{FDR} = E(R_0/\mathrm{R} \mid R > 0),$$

where R_0 is the number of false discoveries and R is the number of results declared significant. In addition, the cost per study sample is related to the number of total reads generated for that sample. The higher the number of reads, the greater the chance of detecting low expression genes. Given a fixed number of subjects, the highest power will be achieved if these subjects are used to sequence with the greatest read depth possible. Thus, the number of subjects and sequence depth are key aspects in power calculation.

As researchers try to understand these data with experiments, paying close attention to the experimental design and the number of biological replicates will be essential in order to have reproducible results in their study. This decision depends on the amount of power

***Corresponding author:** Chung-I Li, Department of Applied Mathematics, National Chiayi University, Chiayi, Taiwan 60004, ROC, Tel: +886-5-2717864; E-mail: chunyilee1@gmail.com

Received May 22, 2014; **Accepted** June 26, 2014; **Published** June 30, 2014

Citation: Shyr D, Li CI (2014) Sample Size Calculation of RNA-sequencing Experiment-A Simulation-Based Approach of TCGA Data. J Biomet Biostat 5: 198. doi:10.4172/2155-6180.1000198

that the researcher hopes to achieve in his or her experiment [8]. Typically, researchers try to aim for a power around 80%. While it would be ideal to have as many subjects as possible to ensure the quality and reproducibility of the results, costs must also be considered.

Material and Methods

Past methods for calculating RNA-seq sample size

Methods of calculating sample size for RNA-seq gene differential expression experiments have been and are being developed. Unlike data sets like microarray that have continuous data, RNA-seq has count data and a skewed distribution. One of the distributions that have been used to model RNA-seq is the Poisson distribution. In [10], sample size formulas based on likelihood ratio test and score test were derived and the procedure of calculating sample size while controlling the false discovery rate (FDR) based on the Poisson distribution was developed. While Poisson may seem to be an appropriate model, the issue of the distribution lies with its critical assumption that the mean and variance must be equal. This assumption has proven to be problematic due to RNA-seq's over-dispersion (variance greater than mean); thus, the Poisson model for RNA-seq has the risk of underestimating the needed sample size, causing the study to be underpowered [10]. An alternative distribution has also been presented: negative binomial. Unlike Poisson, a special case of negative binomial, this distribution can not only model count data, but also have unequal mean and variance, allowing for over-dispersion. In [9], the paper's comparison between the Poisson and negative binomial distribution for the Transcript Regulation data set, which had significant over-dispersion, showed that the latter required a larger sample size than the former. This difference appeared to be more significant as the fold change increased, which, as a result, may signify negative binomial's flaw in overpowering an experiment's sample size. Other analytical methods for estimating RNA-seq sample size have also been developed. For example, [11] derived an explicit sample size formula by using the score test under generalized linear model framework. In this paper, we evaluated the sample size estimations of [10] and [9] by developing a simulation-based approach. Because our method is an empirical approach, we are not limited by any assumptions that the Poisson and negative binomial distribution require. Thus, our method can easily accommodate various RNA-seq data structure based on the data's over-dispersion and fold change.

Power simulation

While closed-form equations for predicting sample size based on the Poisson and negative binomial distribution exist for gene differential expression in RNA-seq, neither distribution can guarantee that their calculated sample size is absolutely correct. Thus, the focus of the study is to develop a simulation-based approach that calculates the power of RNA-seq experiments and estimates the needed sample size and apply the simulation to several TCGA data sets. This approach, known as power simulations, usually follows a series of steps. First, a distribution of parameters, such as sequencing depth and fold change, must be established from some data set that could be from published literature or study. From that data, estimates of the model, including the mean, variance-covariance matrix, and other parameters, can be obtained to help calculate the power. Second, a count data needs to be randomly generated from the distribution with the parameters estimated from step one. Finally, the count data is used to determine whether the sample has sufficient evidence to reject the null hypothesis and be statistically significant [12,13]. Once this is done for each sample, the power of the experiment can be calculated for that particular sample size.

Data sets

Launched in 2006 with funding from the National Cancer Institute and National Human Genome Research Institute, the TCGA was created so that research teams around the world could pool their distinct project results together for public access. With the goal to explore the genomic changes in human cancers, TCGA currently holds more than 20 sequenced tumor types and gives researchers the opportunity to make important discoveries from the sequenced data. Because TCGA continuously collects and characterizes various tumor types from various resources and has a strong infrastructure in pooling cancer genome results from around the world [14-16], choosing data sets from here would have the potential of showing an accurate estimate of each cancer site's estimated power and sample size. In our study, RNA-seq data of three cancer organ sites–lung (LUSC), colorectal (COAD), and breast (BRCA)–were chosen from the TCGA, containing 459, 411, and 1026 samples, respectively, and applied to our power simulation. Our simulation focused on the sequencing data's number of reads; therefore, we only downloaded level 3 data. The data were organized into a matrix with rows representing genes and columns as samples.

Method

All the simulations were conducted with R version 3.0.2. In order to create a simulation that would imitate a gene differential expression of RNA-seq experiment, the following steps were taken in our approach.

A function (Figure 1) was created so that one can input the sample size, RNA-seq data, group values (e.g. 0 and 1, representing no tumor and tumor, control and treatment, etc.), minimum number of reads, FDR cutoff, fold change boundaries, and the number of random samples. The RNA-seq data must be organized into the format that was described above in order for the code to work. Next, the number of genes are stored and the mean count value of each gene for the control is calculated. The genes that have a count value greater than the minimum set for the function are then selected. The edgeR package is then used to analyze the expression values of the selected genes from the RNA-seq data by returning the dispersion values and applying the exact test in order to calculate the fold change values of the samples. After organizing the fold change values based on the set boundaries and randomly selecting them based on the number of genes at the site, the desired sample size, mean, dispersion, and fold change values are used in a loop to create a list of important values. The SimCount function is implemented and produces raw counts using the negative binomial distribution based on the input parameters of the loop. These count values are arranged based on the control and treatment groups and then input into particular edgeR functions, which output the p-adjusted values. Based on the FDR cutoffs and the group of the samples, the false positives, true negatives, true positives, and false negatives are calculated and stored into a matrix. The function, at the end, outputs a list containing the matrix, fold change, dispersion and the number of genes.

To calculate the power from the simulation and display the simulation results with graphs and csv files, another function was created to perform these tasks. Our function provides two different methods of calculating power. The first method uses the sensitivity formula, which is the number of true positives divided by the sum of the number of true positives and false negatives. The second method takes the number of true positives divided by the total number of genes. Both types of powers are compiled with corresponding sample size and the run time as a csv file. The function then produces a scatter plot of the sample size vs. power (sensitivity method) as a pdf file. The

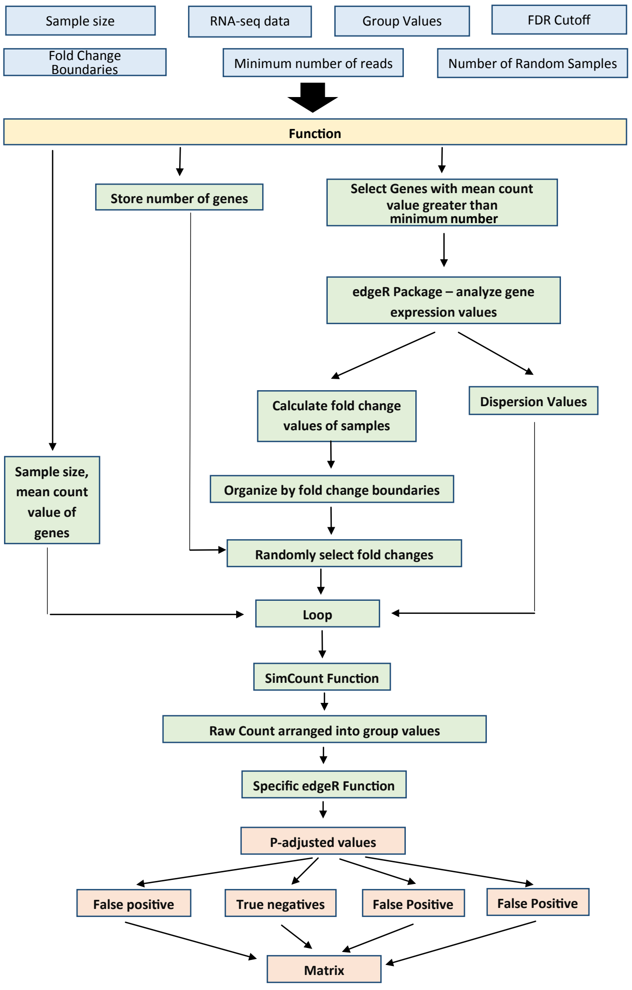
Sample size
RNA-seq data
Group Values
FDR Cutoff
Fold Change Boundaries
Minimum number of reads
Number of Random Samples
Function
Store number of genes
Select Genes with mean count value greater than minimum number
edgeR Package – analyze gene expression values
Calculate fold change values of samples
Dispersion Values
Sample size, mean count value of genes
Organize by fold change boundaries
Randomly select fold changes
Loop
SimCount Function
Raw Count arranged into group values
Specific edgeR Function
P-adjusted values
False positive
True negatives
False Positive
False Positive
Matrix

Figure 1: Flow Chart of Function.

dispersion and fold change of the data are also saved as csv files and the violin plot is provided to display the dispersion results. Because the violin plot is a combination of a box plot and a kernel density plot, it is able to capture the details of the dispersion.

Results

TCGA data set

Simulations were conducted for the three cancer sites of the RNA-seq data from TCGA and the plots of the sample sizes and powers were produced. Figure 2 shows the violin plots of dispersion for three cancer sites. Similar to what other papers, such as [10] and [9], have mentioned, all three sites of the RNA-seq data have similar dispersion distributions that were heavily skewed to the right. From Table 1, it is interesting to note that the dispersion value was between 2 and 2.5 for all three cancer sites at the 95th percentile while the maximum dispersion ranges from 9.686 to 15.88. Therefore, there were relatively few samples in these three cancer sites that had a large dispersion. A simulation was conducted with a FDR of 0.05 and minimum read of 5 and a graph, Figure 3, showing the samples size and power relationship or each cancer site was made when the desired minimum fold change was 2.0. From the results, we found that at 80% power LUSC required a sample size of approximately 18, COAD 20, and BRCA 25, respectively. For COAD, the power values reached a plateau of about 85% as the sample size increased to 38 and greater; on the other hand, LUSC and BRCA reached the highest power of 90% as the sample size increased to 70 and 85, respectively. These sample size results are appropriate given that the variances for BRCA and LUSC are greater than COAD. Note that power should tend to 100% as the sample size increases. Although there is a slow increasing trend in power in Figure 3, it is expected that the power will tend to 100% when the sample size is

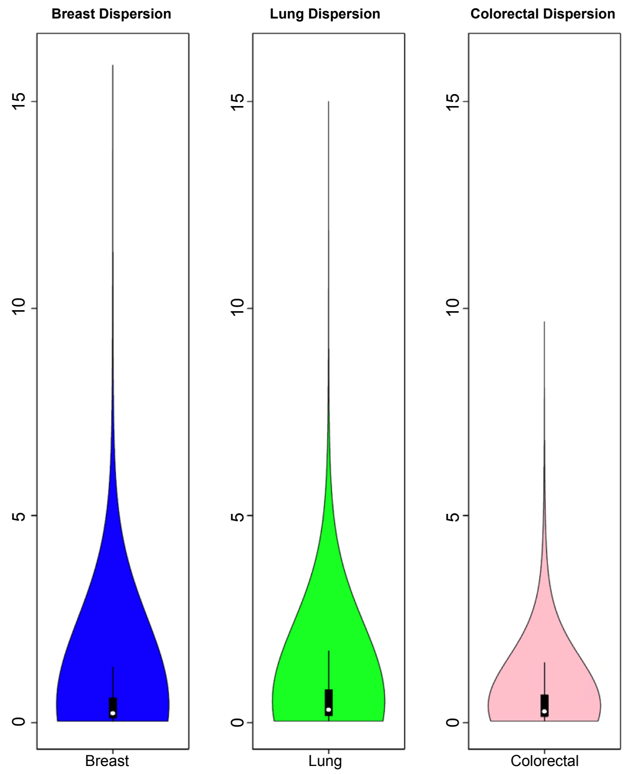

Figure 2: Violin plots of dispersion for three cancer sites.

Summary statistics	*Breast*	*Lung*	*Colorectal*
Min.	0.0329	0.0393	0.0384
1st Qu.	0.1143	0.1657	0.1534
Median	0.2283	0.3141	0.2690
Mean	0.5770	0.6720	0.5707
3rd Qu.	0.6036	0.7914	0.6724
90th Percentile	1.3376	1.6607	1.3910
95th Percentile	2.1355	2.4612	2.0697
Max	15.8800	15.0000	9.6860

Table 1: Summary statistics of dispersion for three cancer sites.

		Power=80%			Power=85%		
		FDR			FDR		
*m	Cancer sites	0.01	0.05	0.1	0.01	0.05	0.1
5	LUSC	26	18	15	37	25	24
	COAD	28	20	17	49	38	35
	BRCA	34	25	20	45	35	30
10	LUSC	25	19	15	40	30	25
	COAD	29	22	18	65	50	40
	BRCA	31	25	20	49	35	33

*m=the minimum number of reads

Table 2: Power vs. Cancer Site Sample Size Estimation.

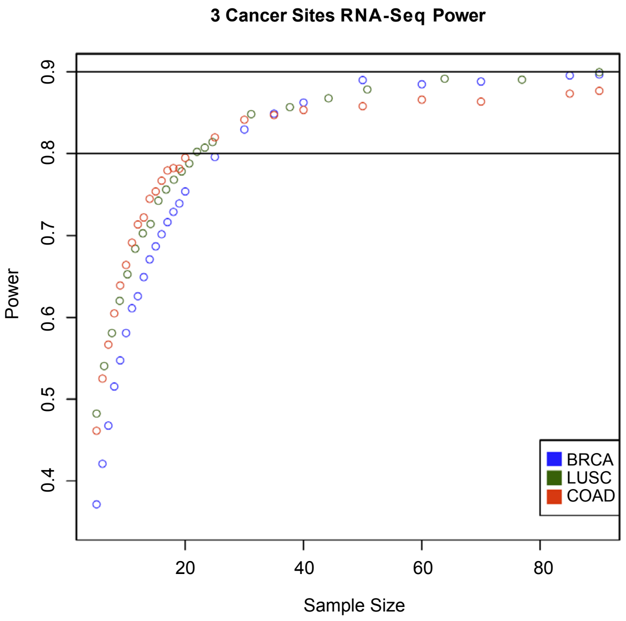

Figure 3: Scatterplot of the samples size and power (minimum reads=5 and FDR=0.05).

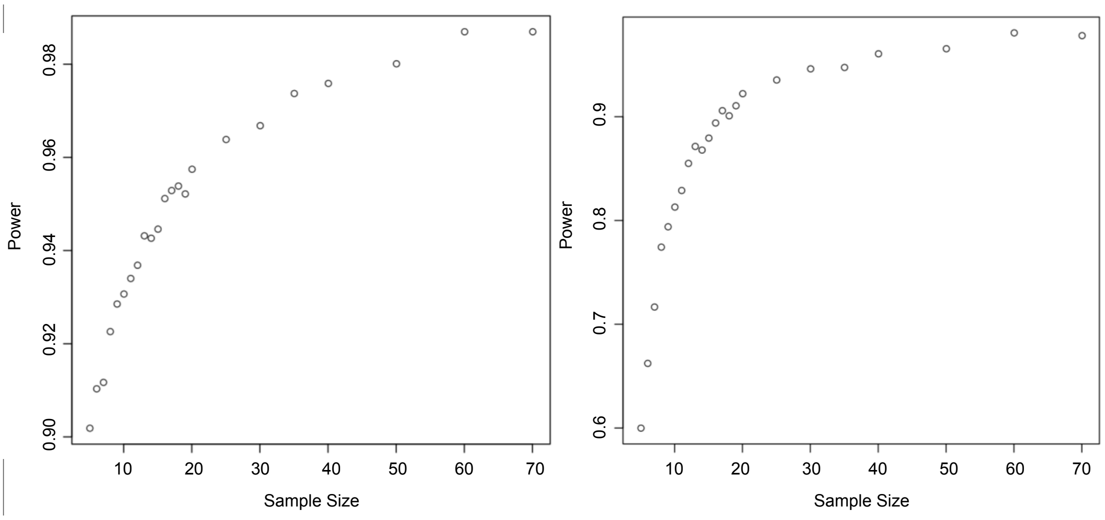

Figure 4:(a) The plot of sample size and power for kidney data set (b) The plot of sample size and power for transcript regulation data set.

sufficiently large. We also ran simulations with different parameters for FDR and minimum reads and found the sample size values, which can be found in Table 2. From these various parameters, COAD was the only cancer site that could not reach 90% power, even at a sample size of 150. The sample size estimation results between minimum reads of 5 and 10 were quite similar, with the latter requiring a few more subjects occasionally as expected.

Kidney data set and transcript regulation data set

[9] considered *kidney data set and transcript regulation data set* as pilot data to test the performance of their method. Here we used this dataset to test our simulation-based method and calculated the sample size under the same settings as [9]. Figure 4a and 4b show the plot of sample size and power for the kidney and transcript regulation data when the desired minimum fold change was 2.0. From [9], their study showed that the kidney dataset required about 15 samples to attain a power of 80% based on the Poisson and negative binomial model. From Figure 4a, a sample size of 15 reached at least a power of 90%, which indicates our method has a higher chance of detecting the true positives at a reduced cost of experimental design. For the transcript regulation dataset, [9] required a sample size of 79 and 31 for the negative binomial model and Poisson model, respectively. Using our method, Figure 4b showed that the same sample sizes of 79 and 31 reached a power of 95% and 90%, respectively. These results indicated that the required sample size based on our method was smaller than Poisson and negative binomial models, which was as expected. While

[9] and [10] provided a conservative estimate of the required sample size, our method is an empirical approach which integrates all the information from the data.

Discussion

In this study, we developed a simulation-based approach to estimate the sample size for RNA-seq gene differential expression experiments. Unlike past sample size estimation methods that have relied on mathematical formulas and distributions which require many assumptions, our empirical approach can accommodate the structure of data based on the data's over-dispersion. Thus, it is recommended to conduct a pilot or feasibility study to generate the preliminary data for sample size calculation if there is no similar existing data that can be used. However, the preliminary data from the pilot study may not be available. In such a situation, we suggest that the parameters of distributions of over-dispersion can be estimated based on the researcher's prior knowledge. From the TCGA data sets, sample size estimations varied among the three cancer sites because of the differences in dispersion and fold change values. Because BRCA had the largest dispersion compared to COAD and LUSC, the number of biological samples required at a power of 80% were greater than the other sites. This remained true even when the FDR and minimum number of reads were adjusted. Our results also showed that each of these three cancer sites had its own unique dispersion distribution, causing the sample size estimation to vary accordingly. Reasons for why COAD could not reach a power of 90% remain uncertain, although its low number of samples in TCGA could be an explanation.

When researchers construct an experimental design, it's important to have preliminary data on the number of biological replicates needed for their experiment. While researchers criticize power analyses for having too many mathematical assumptions, our method overcomes this issue and simply requires RNA-seq data for power and sample size estimation. The flexibility of our method also allows users to modify the proposed procedure of the simulation by using packages other than edgeR, such as DeSeq2 [17], baySeq [18], ShrinkSeq [19], NBPSeq [20], and SAMseq [21], for calculating power or sample size. From our simulation-based approach, researchers will not only have a better idea in designing their experiments, but also have more faith that their findings accurately represent the story behind their data. To facilitate implementation of sample size calculation, R code is available from the corresponding author.

Acknowledgement

This work was supported by National Cancer Institute grants U01 CA163056, P30 CA068485, P50 CA098131, and P50 CA090949.

References

1. Shyr D, Liu Q (2013) Next generation sequencing in cancer research and clinical application. Biol Proced Online 15: 4.
2. Metzker ML (2010) Sequencing technologies-the next generation. Nat Rev Genet 11: 31-46.
3. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5: 19-21.
4. Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, et al. (2012) Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. N Engl J Med 366: 707-714.
5. Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. Nature 483: 531-533.
6. Problems with scientific research - How science goes wrong. The Economist, 2013.
7. Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 2: 93-113.
8. Chow SC, Wang H, Shao J (2003) Sample Size Calculations in Clinical Research. Chapman & Hall/CRC Biostatistics Series p. 14-100.
9. Li CI, Su PF, Shyr Y (2013) Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. BMC Bioinformatics 14: 357.
10. Li CI, Su PF, Guo Y, Shyr Y (2013) Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution. Int J Comput Biol Drug Des 6: 358-375.
11. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP (2013) Calculating sample size estimates for RNA sequencing data. J Comput Biol 20: 970-978.
12. Psioda M (2012) Random Effects Simulation for Sample Size Calculations Using SAS 1-11.
13. Zhao W, Li AX (2011) Estimating Sample Size through Simulations 1-6.
14. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, et al. (2010) International network of cancer genome projects. Nature 464: 993-998.
15. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061-1068.
16. http://cancergenome.nih.gov/abouttcga/overview
17. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Bioarchive 1-34.
18. Hardcastle TJ, Kelly KA (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11: 422.
19. Van De Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW, et al. (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics 14: 113-128.
20. Di Y, Schafer DW, Cumbie JS, Chang JH (2011) The NBP negative binomial model for assessing differential gene expression from RNA-seq. Statistical Applications in Genetics and Molecular Biology 10: 1-28.
21. Li J, Tibshirani R (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22: 519-536.