A hybrid method for differentially expressed genes identification and ranking from RNA-Seq data

Mohammad Samir Farooqi*

Centre for Agricultural Bioinformatics,
Indian Agricultural Statistics Research Institute,
Library Avenue, Pusa,
New Delhi, 110012, Indıa
Email: samirfarooqi8@gmail.com
*Corresponding author

Devendra Kumar

Department of Statistics,
Central University Haryana,
Jant-Pali, Mahendergarh District, Pali,
Haryana, 123031, Indıa
Email: devendrastats@gmail.com

Dwijesh Chandra Mishra and Anil Rai

Centre for Agricultural Bioinformatics,
Indian Agricultural Statistics Research Institute,
Library Avenue, Pusa,
New Delhi, 110012, Indıa
Email: dwij.mishra@gmail.com
Email: anilrai64@gmail.com

Niraj Kumar Singh

Department of Statistics,
AIAS, Amity University,
Noida, UP, 201313, India
Email: nksingh@amity.edu

Abstract: RNA-Seq has gained immense popularity and emerged as a potential high-throughput platform for identification of differentially expressed (DE) genes. In order to estimate the nature of differential genes, it is important to find statistical distributional property of the data. In the present study we propose a new hybrid model (NBPFCROS) based on parametric and non-parametric statistic for the identification of DE genes. The NBP model based on Compound mixture of Poisson–gamma distribution is used as a parametric statistic and Fold change value derived using fold change rank ordering statistics (FCROS) algorithm is used as non-parametric statistic, we used a gene sıgnificance score pi-value by combining expression fold change (f value) and statistical significance (p-value). The performance of

NBPFCROS model was compared with NBP, FCROS, edgeR and DESeq2 models using synthetic and real RNA-Seq datasets and it was found that the developed model NBPFCROS is more robust as compared to the other models.

Keywords: RNA-Seq; differentially expressed genes; parametric and non-parametric statistic; order statistics; fold change; gene significance score; classification accuracy; gene ranking.

Reference to this paper should be made as follows: Farooqi, M.S., Kumar, D., Mishra, D.C., Rai, A. and Singh, N.K. (2021) ‘A hybrid method for differentially expressed genes identification and ranking from RNA-Seq data’, *Int. J. Bioinformatics Research and Applications*, Vol. 17, No. 1, pp.38–52.

Biographical notes: Mohammad Samir Farooqi is a Scientist working at Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India (http://cabgrid.res.in/cabin/msfarooqi.aspx). He did his MSc in Statistics from Aligarh Muslim University, Aligarh, India. Currently, he is pursuing PhD from Amity Institute of Applied Sciences, Amity University, Noida. He has more than 19 years of experience in research and teaching. He has several research papers in national and international journals of repute and also serving as faculty of Bioinformatics and Computer Applications at PG School, ICAR-IARI, New Delhi. His current area of interests includes bioinformatics, order statistics and generalised order statistics, statistical data analysis, data warehouse and data mining.

Devendra Kumar has done his PhD in Statistics from Aligarh Muslim University, Aligarh and presently working as an Assistant Professor and Head, Department of Statistics, Central University Haryana, Mahendergarh. He has published several research papers in national and international journal of repute. His areas of specialisation are order statistics and generalised order statistics, statistical modelling and statistical inference.

Dwijesh Chandra Mishra is a scientist at Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India (http://cabgrid.res.in/cabin/dcmishra.aspx). He did his PhD in Agricultural Statistics from ICAR-Indian Agricultural Research Institute, New Delhi, India. He has more than 13 years of experience in research. He has several research papers in national and international journals of repute and also serving as faculty of Bioinformatics at ICAR-IASRI, New Delhi. His current areas of interests include computational biology, genome assembly, genomic data warehouse, transcriptomics, system biology, genomic selection and GWAS.

Anil Rai is a Principal Scientist and Head at Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India (http://cabgrid.res.in/cabin/arai.aspx). He did his PhD in Agricultural Statistics from ICAR-Indian Agricultural Research Institute, New Delhi, India. He has more than 25 years of experience in research and teaching. He has several research papers in national and international journals of repute and also serving as faculty of Bioinformatics and Agricultural Statistics at PG School, ICAR-IARI, New Delhi. His current area of interests includes bioinformatics, computational biology, spatial modelling and simulation, data warehousing and data mining, complex survey data analysis.

Niraj Kumar Singh has done his PhD in Statistics from Banaras Hindu University, Varanasi and presently working as an Assistant Professor in Amity University, Noida. He has published 17 research papers in national and international journal of repute and three book chapters. His area of specialisation is statistical modelling, applied statistics and statistical demography.

This paper is revised and expanded version of paper entitled 'Gene expression analysis for Rna-seq data using order statistic based mixture model' presented at *71st Annual Conference of Indian Society of Agricultural Statistics (ISAS), ICAR-DRMR*, Bharatpur, Rajasthan, India, 25–27 November, 2017.

1 Introduction

A transcriptome is a collection of all the transcripts in a genome that includes both protein coding mRNAs and noncoding RNAs. Understanding transcriptomes is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and to understand development of a disease (Wang et al., 2009). In transcriptomic data analysis it is important to accurately quantify the abundance of each transcript within different cells and tissues at different time points and to correlate the changes in gene expression with different time points or conditions. Gene expression is an outcome that uses genetic instructions to produce gene products called proteins that perform essential functions as enzymes, hormones etc. Pathogenic infection from microorganisms can be obtained by measuring the level of gene expression in a cell, tissue or organism. Gene expression can provide valuable information such as infection from microorganisms, determine susceptibility to a particular disease or test whether a particular organism is resistant to specific drug. RNA sequencing (RNA-Seq) and microarrays are the main techniques for measuring gene expression and its regulation. Microarrays analysis was most preferred method during 1990s, it requires knowledge of the target sequences to construct the probe sets, therefore expression measures cannot extend beyond this probe set. Also it generates continuous measurement of expression levels, but the RNA-Seq method has a low background noise with high resolution which allows for a single base resolution and generates digital gene expression counts. Hence, over the period of time, microarrays became obsolete and were replaced by RNA-Seq for the discovery of novel transcripts, differential expression (DE) analysis, splice variant detection etc, with improved accuracy and sensitivity (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2008; Wang et al., 2009). In RNA-Seq analysis, the read counts data is represented as a matrix, with rows representing genes and columns representing samples from one or more populations. Here the main objective is to detect differentially expressed genes under different environmental conditions for a given trait.

Data pre-processing, statistical analysis and functional interpretation are three major steps of analysis for both microarray and RNA-Seq. Pre-processing of microarray data normally includes background correction, normalisation and summation, while pre-processing of RNA-Seq data includes artefact filtering and short read alignment/assembly. Normalisation and filtering are the common method to reduce data variability and data dimensionality. Many of the methods used for microarray data analysis, including the method of identifying genes with fold changes are known to be unreliable because in such methods the statistical variability of the data is not properly addressed. Methods for detecting differential expression in microarray data are well-established but generally not applicable to RNA-Seq data. Recently, several additional methods have been developed specifically for RNA-Seq datasets however some of the statistical methods developed for microarray data analysis can also be applied to RNA-Seq data with or without modifications. From a technological perspective, results obtained from next generation sequencing (NGS) technologies and microarrays agree strongly (Cloonan et al., 2008; Mortazavi et al., 2008; Sultan et al., 2008; Fu et al., 2009; Bradford et al., 2010). However the difficulties associated with the generation of data through NGS technique poses an in-built challenge to accurate analysis and interpretation of DE genes. RNA-Seq data analysis requires a number of issues to be taken care of which includes, biases introduced during library preparation, biases of abundance measures due to the effects of nucleotide composition and the varying length of genes or transcripts, the difference in total number of mapped reads for different samples. Thus the observed mapped read counts cannot be directly compared between samples. In order to handle these issues use of statistical hypothesis tests are required to model RNA-Seq count data for detection of significant DE genes across samples. There have been a number of statistical approaches proposed for differential expression analysis of RNA-Seq data, and they are broadly classified into two categories: parametric and nonparametric.

Initial parametric approaches included technical replicates only, the distribution of feature counts across technical replicates was reported to fit well to a Poisson distribution where the variance was equal to the mean (Marioni et al., 2008), it was found that the Poisson distribution underestimates the variation seen in the data. The frequencies of RNA-Seq reads cannot be adequately modelled by the most commonly used distributions such as normal, binomial or Poisson (Di et al., 2011). Biological replicates are more variable than technical replicates (McIntyre et al., 2011). In RNA-Seq analysis the number of genes is far greater than the number of samples. Another dimension often present in RNA-Seq datasets is the number of replicates. Since the RNA-Seq protocol is highly reproducible technical replicates are usually not necessary, and instead 2 to 3 biological replicates are used to reduce the degree of noise resulting from biological variations (Fang et al., 2012). Negative Binomial (NB) distribution because of its ability to deal with the over dispersion problem was proposed. Several authors (Robinson and Smyth, 2007; Anders and Huber, 2010; Hardcastle and Kelly, 2010) have proposed differential expression methods based on the negative binomial distribution. For two-group comparisons, the NB distribution permits an exact test (Robinson and Smyth, 2007). Different tools, such as DEGSeq (Wang et al., 2010), DEseq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), baySeq (Hardcastle and Kelly, 2010), PoissonSeq (Li et al., 2012) and gfold (Feng et al., 2012) etc, have been developed to enable differential expression analysis of RNA-Seq data. However, the results obtained from these tools are usually different depending on the differential expression algorithms used. If the distributional assumption holds, these parametric approaches are generally very efficient and reliable. But, violation of distributional assumptions or a poor estimation of parameters often leads to unreliable results. Therefore models based on non-parametric methods have also been developed to counter these problems. It has been shown that the Fold Change based selection of genes gives better results regardless of the technology being used (Guo et al., 2006; Shi et al., 2005; Chen et al., 2007), NOISeq (Tarazona et al., 2011) is a data-adaptive nonparametric approach that uses both log fold changes and absolute expression differences as test statistics. Li and Tibshirani (2013) proposed that existing methods based on Poisson or negative-binomial models are useful but can be heavily influenced by 'outliers' in the data, they introduced Samseq a simple, non-parametric method with resampling which also accounts for the difference in sequencing depths. This method utilises a Wilcoxon statistic and was more robust than parametric methods in such situations. LFCseq (Lin et al., 2014) proposed a new data-driven nonparametric approach for differential expression analysis of RNA-Seq data. It was based on a similar principle to NOISeq but uses only log fold changes as the test statistic. DE analysis of RNA-Seq data is still developing and new methods are continuously being introduced but to date, there is no general consensus that, which method performs best in a given situation. Given the limitation of small sample sizes in RNA-Seq experiments, robust test procedures which safeguard against the departure of model assumptions are necessary (Fang et al., 2012). Concepts of compound distribution (Di et al., 2011; Anjum et al., 2016) have also been applied to account for the variability from different sources of variation, for identification of differentially expressed genes. In this paper we discuss a new approach for the development of a hybrid model, based on parametric and nonparametric statistic for the identification of DE genes. This hybrid model accounts for the extra variation in the analysis of sequence count data and derive a score on the basis of which the identified DE genes are ranked.

2 Materials and methods

First step in the development of hybrid model for gene ranking was to select an appropriate model from parametric and non-parametric statistic and combined them using the approach given by Becker (1994). For this the negative binomial power (NBP) model given by Di et al. (2011) was selected from the parametric statistic and Fold change rank ordering statistics (FCROS) given by Dembele and Kastner (2014) was selected from the nonparametric statistic. In a parametric approach for an RNA-Seq dataset, the expression level of a specific gene, say $Y_{ij}^{(k)}$, is the total number of short sequences which gets aligned to the *i*th gene in *j*th replication and *k*th sample. Di et al. (2011) applied a generalised negative binomial distribution, known as the negative binomial power (NBP) distribution, to test for differential expression. The NBP distribution is a gamma mixture of Poisson distributions; if $Y|Z \sim POI(Z)$ and $Z \sim \Gamma$ with mean μ and variance $\phi\mu^{\alpha}$, then marginal distribution of Y is NBP, by assuming NBP as distributed read counts, $\text{Var}(Y_{ij}^{(k)}) = \mu_i^{(k)}(1+\phi(\mu_i^{(k)})^{\alpha-1}$. NBP method extends an exact test proposed by Robinson and Smyth (2007, 2008) and allows for flexible mean and variance relationship. Here the dispersion parameter Φ is common to all genes; the mean-variance relationship is given flexibility via the power parameter α. The NBP test is constructed as an exact test based on the NBP assumption. The null hypothesis for NBP is $\lambda_{1g} = \lambda_{2g}$. Where λ_{1g} and λ_{2g} are expression of *g*th gene in 1st and 2nd samples, respectively. The probability of statistical significance (*p*-value) for each gene is obtained using this model.

The fold change (FC) which is calculated as a ratio of averages from control and test sample values was initially used by (Schena et al., 1995; Lockhart et al., 1996). FC is a basic and widely used measure for identifying differential gene expression; however, the raw fold change is unreliable as it does not take into account the uncertainty of gene expression measures under the two conditions being compared. Thus, other statistical methods were introduced. Researchers have also attempted to combine the fold change and *p*-value to provide more meaningful results by setting cutoffs for both the fold change and *p*-value (Cui and Churchill, 2003; Xiao et al., 2014). It has been shown that the FC based selection of genes leads to more reproducible results irrespective of the technology that is used (Guo et al., 2006; Shi et al., 2005; Chen et al., 2007). Dembele and Kastner (2014) gave a new FC-based method Fold change rank ordering statistics (FCROS) and showed that it is powerful to detect DE genes in noisy datasets. This method assigns a ranking statistic to DE genes. Let there be expression values obtained for *y* genes in *x*1 control and *x*2 test samples, then pairwise comparisons for $k \leq x1x2$ are performed and FCs for each gene (test/control) is computed. In each comparison, the y FCs obtained are sorted in increasing order and their corresponding ranks are associated to genes. Hence, for gene *i*, we get a vector $ri = (r_{i1}\ r_{i2}\ \ldots\ r_{ij}\ \ldots\ , r_{ik})$ where r_{ij} corresponds to the rank of the FC for gene *i* in the *j* comparison ($j = 1, \ldots, k$). Robust average of rank for each gene ($i = 1, 2, \ldots, y$) using its *k* values was calculated and FCROS algorithm (Dembele and Kastner, 2014), was applied to obtain *f*-values which are the probabilities associated to fold change ranks ordering statistics. An *f*-value close to 0.5 corresponds to an equally expressed (EE) gene, while down- and up-regulated genes have *f*-values close to 0 and 1, respectively.

Following, Xiao et al. (2014) who provided a new ranking method for genes based on combination of expression change (*f*-value) and statistical significance (*p*-value) for microarray data. A new hybrid model (NBPFCROS) for ranking genes based on non-parametric statistics *f*-value derived through FCROS model and the statistical significance p value obtained through parametric NBP model for RNA-Seq data has been proposed as:

$$\pi_i = -\left(\log \frac{f_i}{1-f_i} + \log \frac{P_i}{1-P_i} \right) \Big/ c$$

where

$$c = \sqrt{4\pi^2 / 7}$$

f_i: scaled fold rank order statistics from FCROS method for *i*th gene

p_i: *p*-value derived through NBP method for *i*th gene.

To bring *f*-value and *p*-value on same scale following transformation was made.

$$f_i = 2 \times (\text{abs}(f_i.\text{value-}0.5))$$

Pre-processing or normalisation of the dataset has been done by random sampling of counts to make the effective library sizes equal (column sums of the count matrix multiplied by normalisation factors) as suggested by Robinson and Oshlack (2010). The NBPFCROS model gives a score π_i for the *i*th gene, *R* package Metap (Michael, 2016) and logitp method (Becker, 1994), was used to combine the p and f values obtained from NBP and FCROS method, respectively.

Further, empirical *P* values corresponding to each π_i score was obtained by generating bootstrap samples (Davison and Hinkley, 1997; Angelo and Brian, 2016), on the basis of which the genes are ranked in the increasing order of the *p* values.

$$P_{\text{value}}(\pi_i) = \frac{1}{N} \sum_{j=1}^{N} I(\pi_{ij} \geq \pi_i)$$

where

J: 1, 2, …, *N* (No. of bootstrap samples)

π_{ij}: π score for *i*th gene in *j*th bootstrap sample

π_i: π score for *i*th gene in original sample.

R codes were written for the entire process of generating combined π-values by obtaining the *p* and *f* values for each gene from their respective models NBP and FCROS, and ranking of genes on the basis of *p* values corresponding to the combined score π_i.

3 Performance evaluation

The performance of the developed hybrid model along with FCROS and NBP models (given in Table 3(a)–(c) was compared on the basis of classification. The performance criteria i.e., mean classification accuracy (CA) was computed by using a sliding window size technique. Here, the window sizes refer to the number of ranked genes obtained by using each gene selection method. Moreover, the window sizes were taken as 50, 100, 150, …, 950, 1000 with a sliding length of 50. Further, the top ranked genes, selected by gene selection method, were then used in SVM classifiers with linear kernel to predict the classes of the samples (stress; +1/ control; –1) on three different datasets (one real and two simulated datasets). The CA was computed by training the SVM classifiers for each sliding window sizes over 5 fold cross validation.

Further *R* codes were also written for generating simulated dataset and testing model accuracy and classification. For all this purpose *R* version 3.2.3 was used and the library of other *R* packages ‘boot’(), ‘e1071’(), ‘metap’(), ‘xlsx’(), ‘NBPseq’(),‘fcros’(), ‘edgeR’(), and ‘DESeq2’ () were utilised for comparing the models and generating result.

4 Results and discussion

In order to obtain the number of differentially expressed genes through the developed hybrid model and compare the number of DE genes derived through the other models we used Arabidopsis thaliana dataset, accessed from NBPseq package of R (Di et al., 2011). The dataset contains 26,222 by six matrix of RNA-Seq read frequencies. The matrix contains the frequencies of RNA-Seq reads mapped to genes in a reference database. Rows correspond to genes and columns correspond to independent biological samples. The dataset was run on all the above mentioned models and the number of differentially expressed genes for all the cases were obtained (Table 1). The screening parameters for obtaining differentially expressed up regulated and down regulated genes were kept unchanged for parametric (NBP) and nonparametric (FCROS) method, as suggested by Di et al. (2011) and Dembele and Kastner (2014), respectively. For NBPFCROS, genes with $p < 0.05$ were identified as differentially expressed. We used scaled π scores, to select the up regulated and down regulated genes from these differentially expressed genes. As the distribution of π scores was skewed, median value was used for segregating the up regulated and down regulated genes. Genes with values of $\pi >$ median were taken as up regulated and $\pi <$ median as down regulated genes. From Table 1 we can observe that number of differentially expressed genes obtained in the case of our hybrid model

NBPFCROS is considerably decreased. The performance of NBPFCROS is better in terms of controlling the false discovery rate and able to detect the true DE genes more precisely as compared to the other methods.

Table 1 Number of DE genes obtained through each model

	FCROS	*NBP*	*NBPFCROS*	*DESeq2*	*edgeR*
Number of differentially expressed genes	3680	1842	1302	2160	2326
Number of up regulated genes	1863	961	652	1086	1292
Number of down regulated genes	1818	857	650	1074	1034

To evaluate the performance of our developed model NBPFCROS, we used synthetic and real RNA-Seq datasets. The synthetic dataset following parametric distribution was generated using compcodeR() package (Soneson, 2014). The simulation was performed following the description by Soneson and Delorenzi (2013). The count dataset contained 15,000 genes for two groups of 15 samples each, where 10% of the genes are simulated to be deferentially expressed between the two groups (equally distributed between up- and down regulated in group 2 compared to group 1). Furthermore, the counts for all genes were simulated from a Negative Binomial distribution with the same dispersion in the two sample groups. For simulating dataset following non parametric distribution, we used package SimSeq (Benidt and Nettleton, 2015), the generated count dataset contained 15,000 genes for two groups of 15 samples each, where 10% of the genes were simulated to be deferentially expressed between the two groups.

For real dataset, we used KIRC RNA-seq dataset (The version of the KIRC dataset unc.edu_KIRC.IlluminaHiSeq_RNASeqV2.Level_3.1.5.0 accessed from Simseq package of R) containing 20,531 genes and 72 paired columns of data with rows corresponding to genes and columns corresponding to replicates; replic vector specifies replicates and treatment vector specifies non-tumour and tumour group samples respectively within replicate (The Cancer Genome Atlas Research Network, 2013). First 20 samples from both the groups were included in the dataset for validity check of the developed model. List of top 50 differentially expressed genes, ranked on the basis of *p* values, obtained through hybrid model NBPFCROS is given in Table 2.

Table 2 List of top 50 ranked genes

Gene name	*Pi-value*	*p-value*
ALDOA|226	0.013185	0
ANGPTL4|51129	1.55E-21	0
ATP1A1|476	1.25E-09	0
C3orf71|646450	1.3E-07	0
CD59|966	0.021953	0
CDH11|1009	8.97E-13	0
GANAB|23193	0.004646	0
GNB1L|54584	0.000359	0
HKDC1|80201	2.05E-06	0
IGF2R|3482	2.44E-14	0

Table 2 List of top 50 ranked genes (continued)

Gene name	*Pi-value*	*p-value*
LDB2|9079	7.05E-07	0
NDNL2|56160	0.000152	0
NDUFA3|4696	2.47E-22	0
NMT2|9397	1.3E-17	0
PGF|5228	0.000616	0
PLEKHO2|80301	3E-15	0
RGS3|5998	5.47E-10	0
SERP1|27230	1.03E-08	0
SOCS7|30837	1.85E-06	0
SPAG8|26206	2.09E-08	0
TMPO|7112	4.89E-11	0
TUB|7275	2.03E-06	0
UBASH3B|84959	0.159337	0
VDAC3|7419	1.07E-15	0
VHL|7428	6.45E-11	0
AIF1L|83543	6.37E-16	0.002
ANXA4|307	1.22E-10	0.002
B2M|567	0.006522	0.002
C2orf86|51057	2.52E-16	0.002
CAND2|23066	0.129763	0.002
CD69|969	6.7E-06	0.002
COL27A1|85301	2.92E-11	0.002
COL3A1|1281	7.05E-09	0.002
DCLRE1B|64858	1.55E-16	0.002
EEF1E1|9521	0.473419	0.002
FMOD|2331	4.69E-09	0.002
GLRX3|10539	1.06E-06	0.002
HKR1|284459	1.91E-07	0.002
HLA-A|3105	1.87E-05	0.002
HLA-DMB|3109	1.83E-11	0.002
HLA-DQA2|3118	1.06E-08	0.002
HSCB|150274	5.21E-18	0.002
MEA1|4201	5.47E-14	0.002
PEA15|8682	6.4E-06	0.002
PFKL|5211	7.44E-11	0.002
PLSCR4|57088	2.14E-10	0.002
RPL11|6135	0.001404	0.002
RPL8|6132	0.00222	0.002
TFB2M|64216	3.6E-17	0.002
TUBA4A|7277	0.002591	0.002

We compared the results obtained from NBPFCROS with the NBP, FCROS, edgeR and DESeq2 methods to test for its validity and robustness. Support vector machine (SVM) with linear kernel was used for evaluating the classification accuracy of the model. Table 3(a)–(c) clearly suggest that the NBPFCROS method has better mean classification accuracy rate for the simulated data based on parametric distribution whereas in the case of simulated data based on non-parametric distribution and on real dataset, the classification accuracy rate for NBPFCROS is better than NBP, edgeR and DESeq2 models. Moreover from Table 3(a)–(c), we can also see that 100% classification accuracy rate is achieved with less number of predictors (widow size) for NBPFCROS in comparison to other models. As the hybrid model NBPFCROS has performed well on all the used datasets, it suggest that gene significance score π-value obtained through NBPFCROS is more robust in terms of identifying true DE genes and ranking them according to their significance as compared to the individual NBP and FCROS algorithm.

Table 3(a) Performance of various methods in terms of classification accuracy (CA)

	CA-real data				
Window size	*NBP*	*FCROS*	*NBPFCROS*	*edgeR*	*DESeq2*
50	95.1	97.1	96.1	38.2	37.8
100	96.4	97.1	96.2	37.9	38.8
150	97.0	97.1	96.8	37.0	38.5
200	97.3	97.0	96.6	36.1	36.6
250	97.1	97.1	96.8	39.3	37.6
300	97.8	97.1	96.8	38.3	35.8
350	97.8	97.0	96.6	38.2	36.8
400	97.2	97.3	96.7	39.2	37.8
450	97.5	97.0	96.4	37.4	35.0
500	96.9	97.4	96.4	37.2	38.8
550	97.5	97.5	96.6	37.3	34.3
600	97.5	97.5	96.4	38.2	36.8
650	97.5	97.2	95.6	39.6	36.9
700	97.1	97.3	95.9	37.1	39.2
750	97.5	97.3	96.0	40.5	35.9
800	97.5	97.5	95.7	36.5	37.5
850	97.2	97.5	95.8	38.9	37.2
900	97.5	97.3	96.5	34.4	37.5
950	97.5	97.3	95.9	38.6	35.7
1000	97.5	97.5	96.8	38.3	35.0

Table 3(b) Performance of various methods in terms of classification accuracy (CA)

	CA-simulated data following-parametric distribution				
Window size	*NBP*	*FCROS*	*NBPFCROS*	*edgeR*	*DESeq2*
50	95.2	87.5	98.3	11.3	8.7
100	97.2	88.3	98.9	10.4	9.0
150	96.7	88.4	99.6	9.4	13.0
200	98.0	89.6	99.7	9.8	10.6
250	97.9	90.0	99.9	11.9	5.6
300	98.2	89.7	99.7	10.0	7.5
350	97.4	91.9	99.7	11.4	10.5
400	97.8	90.2	100	13.0	8.3
450	98.6	90.6	100	9.8	11.8
500	98.5	92.7	100	11.5	14.6
550	98.7	91.6	100	7.8	10.0
600	98.8	93.7	100	11.9	10.7
650	98.5	93.3	100	11.5	12.8
700	99.7	94.2	100	9.7	8.3
750	99.7	94.0	100	7.6	10.6
800	100	95.2	100	10.0	8.3
850	100	93.3	100	9.3	11.1
900	100	96.9	100	10.4	14.6
950	100	93.3	100	14.3	9.5
1000	100	97.0	100	11.1	13.9

Table 3(c) Performance of various methods in terms of classification accuracy (CA)

	CA-simulated data following non-parametric distribution				
Window size	*NBP*	*FCROS*	*NBPFCROS*	*edgeR*	*DESeq2*
50	78.4	91.8	83.1	37.2	36.1
100	80.7	89.6	83.6	35.8	34.7
150	83.0	91.4	86.5	37.0	36.1
200	83.9	90.1	88.1	34.5	37.6
250	84.1	91.1	90.0	36.3	36.7
300	86.5	90.7	88.5	34.7	35.3
350	88.8	90.1	89.6	35.6	36.0
400	88.3	89.5	90.6	33.3	34.4
450	90.0	91.6	87.8	36.7	34.7
500	88.8	89.2	90.3	37.1	38.3
550	87.1	91.2	89.8	35.8	36.2
600	87.9	91.3	88.3	34.3	32.9

Table 3(c) Performance of various methods in terms of classification accuracy (CA) (continued)

	CA-simulated data following non-parametric distribution				
Window size	*NBP*	*FCROS*	*NBPFCROS*	*edgeR*	*DESeq2*
650	88.7	92.2	89.2	37.2	35.1
700	86.7	92.7	89.2	37.2	36.7
750	88.2	92.2	90.2	33.6	36.1
800	89.3	93.6	90.0	32.7	32.7
850	88.9	94.6	91.0	35.9	38.9
900	90.4	95.6	88.6	32.5	37.5
950	88.1	94.5	90.0	39.5	36.7
1000	89.4	95.0	88.7	36.7	40.0

5 Conclusion

RNA-Seq method is a widely applied technique in biological research. Efficient and precise statistical method for RNA-Seq data analysis is essential to the advancement of Genomics and Proteomics research. Generating an accurate list of differentially expressed genes is the basis for pathway or gene set enrichment analysis. RNA-Seq data takes the form of counts, so models based on the normal distribution are generally unsuitable. Since current methods for this problem of RNA-Seq data analysis are mostly based on Poisson or negative-binomial models which are useful, but the results may not be reliable if the outliers are present in the data.

All these parametric as well as the non-parametric approaches are developed to increase the power of detection, so that maximum number of true DE genes is identified. However the violation of the model distribution increases the false positive rate and decreases the number of true DE genes. A gene set with a large number of false positives will compromise these analyses. Here we have introduced a new combined method based on parametric and non-parametric distribution (NBPFCROS) which when compared with the individual parametric (NBP) and non-parametric (FCROS) methods, has been found to have increase power of detection in terms of identifying true differentially expressed genes and also consistent classification accuracy across real and simulated datasets.

References

Anders, S. and Huber, W. (2010) ‘Differential expression analysis for sequence count data’, *Genome Biology*, Vol. 11, No. 10, p.R106.

Angelo, C. and Brian, R. (2016) *boot: Bootstrap R (S-Plus) Functions*, R package version 1.3-18.

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A. and Rai, A. (2016) ‘Identification of differentially expressed genes in RNA-Seq data of arabidopsis thaliana: a compound distribution approach’, *Journal of Computational Biology*, Vol. 23, No. 4, pp.1–9.

Becker, B.J. (1994) ‘Combining significance levels’, in Cooper, H. and Hedges, L.V. (Eds.): *A Handbook of Research Synthesis*, Russell Sage, New York, Chapter 15, pp.215–230.

Benidt, S. and Nettleton, D. (2015) ‘SimSeq: a nonparametric approach to simulation of RNA sequence datasets’, *Bioinformatics*, Vol. 31, pp.2131–2140.

Bradford, J.R., Hey, Y., Yates, T., Pepper, S.D. and Mille, C.J. (2010) 'A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling', *BMC Genomics*, Vol. 11, p.282.

Chen, J.J., Hsueh, H.M., Delongchamp, R.R., Lin, C.J. and Tsai, C. (2007) 'A reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data', *BMC Bioinformatics*, Vol. 8, p.412.

Cloonan, N., Forrest., A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J. and Grimmond, S.M. (2008) 'Stem cell transcriptome profiling via massive-scale mRNA sequencing', *Nature Methods*, Vol. 5, pp.613–619.

Cui, X. and Churchill, G.A. (2003) 'Statistical tests for differential expression in cDNA microarray experiments', *Genome Biology*, Vol. 4, No. 4, p.210.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge.

Dembele, D. and Kastner, P. (2014) 'Fold change rank ordering statistics: a new method for detecting differentially expressed genes', *BMC Bioinformatics*, Vol. 15, p.14.

Di, Y., Schafer, D.W., Cumbie, J.S. and Chang, J.H. (2011) 'The NBP negative binomial model for assessing differential gene expression from RNA-Seq', *Statistical Applications in Genetics and Molecular Biology*, Vol. 10, pp.1–28.

Fang, Z., Martin, J. and Wang, Z. (2012) 'Statistical methods for identifying differentially expressed genes in RNA-Seq experiments', *Cell and Bioscience*, Vol. 2, p.26.

Feng, J., Meyer, C.A., Wang, Q., Liu, S.L., Liu, S.X. and Zhang, Y. (2012) 'GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data', *Bioinformatics*, Vol. 28, No. 21, pp.2782–2788.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R. and Khaitovich, P. (2009) ' Estimating accuracy of RNA-Seq and microarrays with proteomics', *BMC Genomics*, Vol. 10, p.161.

Guo, L., Lobenhofer, E.K., Wang, C., Shippy, R., Harris, S.C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F.M., Hurban, P., Phillips, K.L., Xu, J., Deng, X., Sun, Y.A., Tong, W., Dragan, Y.P. and Shi, L. (2006) 'Rat toxicogenomic study reveals analytical consistency across microarray platforms', *Nature Biotechnology*, Vol. 24, No. 9, pp.1162–1169.

Hardcastle, T.J. and Kelly, K.A. (2010) 'baySeq: empirical Bayesian methods for identifying differential expression in sequence count data', *BMC Bioinformatics*, Vol. 11, p.422.

Li, J. and Tibshirani, R. (2013) 'Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data', *Statistical Methods in Medical Research*, Vol. 5, pp.519–536.

Li, J., Witten, D.M., Johnstone, I.M. and Tibshirani, R. (2012) 'Normalization, testing, and false discovery rate estimation for RNA-sequencing data', *Biostatistics (Oxford, England)*, Vol. 13, No. 3, pp.523–538.

Lin, B., Zhang, L. and Chen, X. (2014) 'LFCseq: a nonparametric approach for differential expression analysis of RNA-Seq data', *BMC Genomics*, Vol. 15, No. 10, p.S7.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nature Biotechnology*, Vol. 14, pp.1675–1680.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) 'RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays', *Genome Research*, Vol. 18, pp.1509–1517.

McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J. and Nuzhdin, S.V. (2011) 'RNA-seq: technical variability and sampling', *BMC Genomics*, Vol. 12, p.293.

Michael, D. (2016) *metap: Meta-Analysis of Significance Values*, R package version 0.7.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', Nature. Methods, Vol. 5, pp.621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) 'The transcriptional language of the yeast genome defined by RNA sequencing', *Science*, Vol. 320, No. 5881, pp.1344–1349.

Robinson, M.D. and Oshlack, A. (2010) 'A scaling normalization method for differential expression analysis of RNA-Seq data', *Genome Biology*, Vol. 11, p.R25.

Robinson, M.D. and Smyth, G.K. (2007) 'Moderated statistical tests for assessing differences in tag abundance', *Bioinformatics*, Vol. 23, pp.2881–2887.

Robinson, M.D. and Smyth, G.K. (2008) 'Small-sample estimation of negative binomial dispersion, with applications to Sage data', *Biostatistics*, Vol. 9, pp.321–332.

Robinson, M.D., McCarthy, D.J., Smyth, G.K. and (2010) 'edgeR: a bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, Vol. 26, No. 1, pp.139–140.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) 'Quantitative monitoring of gene expression patterns with a complementary', *Science*, Vol. 270, No. 5235, pp.467–470.

Shi, L., Tong, W., Fang, H., Scherf, U., Han, J., Puri, R.K., Frueh, F.W., Goodsaid, F.M., Guo, L., Su, Z., Han, T., Fuscoe, J.C., Xu, Z.A., Patterson, T.A., Hong, H. and Xie, Q. (2005) 'Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedure are essential', *BMC Bioinformatics*, Vol. 6, No. 2, pp.1471–2105.

Soneson, C. (2014) 'compcodeR – an R package for benchmarking differential expression methods for RNA-Seq data', *Bioinformatics*, Vol. 30, No. 17, pp.2517–2518.

Soneson, C. and Delorenzi, M. (2013) 'A comparison of methods for differential expression analysis of RNA-seq data', *BMC Bioinformatics*, Vol. 14, p.91.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. and Yaspo, M.L. (2008) 'A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome', *Science*, Vol. 321, pp.956–960.

Tarazona, S., Garćıa-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) 'Differential expression in RNA-Seq: a matter of depth', *Genome Research*, Vol. 21, pp.2213–2223.

The Cancer Genome Atlas Research Network (2013) 'Comprehensive molecular characterization of clear cell renal cell carcinoma', *Nature*, Vol. 499, No. 7456, pp.43–49.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, Vol. 456, pp.470–476.

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) 'DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data', *Bioinformatics*, Vol. 26, pp.136–138.

Wang, Z., Gerstein, M. and Synder, M. (2009) 'RNA-Seq: a revolutionary tool for transciptomics', *Nature Review Genetics*, Vol. 10, No. 1, pp.57–63.

Xiao, Y., Hsiao T.H., Suresh, U., Chen, H.I., Wu, X., Wolf, S.E. and Chen, Y. (2014) 'A novel significance score for gene selection and ranking', *Bioinformatics*, Vol. 30, pp.801–807.

Bibliography

Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M. and Caudy, A.A. (2009) 'Measuring differential gene expression by short read sequencing: Quantitative comparison to 2-channel gene expression microarrays', *BMC Genomics*, Vol. 10, p.221.

Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997) 'Ratio-based decisions and the quantitative analysis of cDNA microarray images', *Journal of Biomedical Optics*, Vol. 2, pp.364–374.

Churchill, G.A. (2002) 'Fundamentals of experimental design for cDNA microarrays', *Nature Genetics*, Vol. 32, pp.490–495.

Efron, B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press.

Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2001) 'Microarrays and their use in a comparative experiment', *Journal of the American Statistical Association*, Vol. 96, pp.1151–1160.

Evgenia, D., Kurt, H., Friedrich, L., David, M. and Andreas, W. (2011) *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien, R package version 1.5-25.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) 'Highly integrated single-base resolution maps of the epigenome in Arabidopsis', *Cell*, Vol. 133, pp.523–536.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2', *Genome Biology*, Vol. 15, p.550.

Patterson, T.A., Lobenhofer, E.K., Fulmer-Smentek, S.B., Collins, P.J., Chu, T.M., Bao, W., Fang, H., Kawasaki, E.S., Hager, J., Tikhonova, I.R., Walker, S.J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J.C., Tong, W., Shi, L. and Wolfinger, R.D. (2006) 'Perfomance comparison of one-color and two-color platformswithin the microarray quality control (MAQC) project', *Nature Biotechnology*, Vol. 24, No. 9, pp.1140–1150.

Peart, M.J., Smyth, G.K., Van Laar, R.K., Bowtell, D.D., Richon, V.M., Marks, P.A., Holloway, A.J. and Johnstone, R.W. (2005) 'Identification and functional significance of genes regulated structurally different histone deacetylaseinhibitors', *Proc. Nat. Acad. Sci., U.S.A.*, Vol. 102, No. 10, pp.3697–3702.

Shaik, J. and Yeasin, M. (2007) 'Ranking function based on higher order statistics (RF-HOS) for two-sample microarray experiments, bioinformatics research and applications', *Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, Vol. 4463, pp.97–108.

Yang, Y. and Speed, T. (2002) 'Design issues for cDNA microarray experiments', *Nature Reviews*, Vol. 3, pp.579–588.