G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2017;15(1):51-53
https://doi.org/10.5808/GI.2017.15.1.51

APPLICATION NOTE

TRAPR: R Package for Statistical Analysis and Visualization of RNA-Seq Data

Jae Hyun Lim[§], Soo Youn Lee[§], Ju Han Kim*

Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics and Systems Biomedical Informatics Research Center, Seoul National University College of Medicine, Seoul 110799, Korea

High-throughput transcriptome sequencing, also known as RNA sequencing (RNA-Seq), is a standard technology for measuring gene expression with unprecedented accuracy. Numerous bioconductor packages have been developed for the statistical analysis of RNA-Seq data. However, these tools focus on specific aspects of the data analysis pipeline, and are difficult to appropriately integrate with one another due to their disparate data structures and processing methods. They also lack visualization methods to confirm the integrity of the data and the process. In this paper, we propose an R-based RNA-Seq analysis pipeline called TRAPR, an integrated tool that facilitates the statistical analysis and visualization of RNA-Seq expression data. TRAPR provides various functions for data management, the filtering of low-quality data, normalization, transformation, statistical analysis, data visualization, and result visualization that allow researchers to build customized analysis pipelines.

Keywords: base sequence, gene expression profiling, molecular sequence data, programming languages, sequence analysis/RNA, software

Availability: TRAPR is written in R (the version 2.15), and is available at http://www.snubi.org/software/trapr.

Introduction

High-throughput mRNA sequencing technology has developed at great pace in recent years [1]. Data from RNA sequencing (RNA-Seq) experiments across many species and tissue types are available for free access through public repositories. While RNA-Seq data have a wide range of applications, such as alternative splicing research, fusion gene finding, novel transcript discovery, etc., the most important and widely considered application is the quantification of gene expression profiles and the assessment of differentially expressed genes (DEGs) [2].

Evaluating differential expression in conditions by RNA-Seq is a multi-step process [3]. R/bioconductor [4] has been used to develop tools for the statistical analysis of RNA-Seq data. Some packages stem from classical methods for microarray data analysis, like the t test. Others, like edgeR [5], DESeq [6], DEGSeq [7], and baySeq [8], have recently been developed to the characteristics of RNA-Seq data. However, different packages partially support varying steps of the multi-step process in a very inconsistent manner. Moreover, no R packages support data filtering steps to improve the statistical power and control outliers that might have an undesirable influence on further analysis [9].

This study proposes TRAPR (Total RNA-Seq Analysis Package for R, http://www.snubi.org/software/trapr), an integrated pipeline for the analysis of RNA-Seq gene expression data. TRAPR uses gene expression tables to perform all RNA-Seq analyses, including data preprocessing, filtering, normalization, and statistical tests. TRAPR also provides visualization functions for data exploration and results' summarization. TRAPR provides a unique way of combining state-of-the-art analysis methods in an integrated pipeline for comprehensive RNA-Seq data analysis. For instance, upper-quartile normalization followed by zero-value

Received January 23, 2017; Revised February 7, 2017; Accepted February 15, 2017

*Corresponding author: Tel: +82-2-740-8320, Fax: +82-2-747-8928, E-mail: Juhan@snu.ac.kr

[§]These authors contributed equally to this work.

filtering, variance stabilizing normalization (VSN) [10], and edgeR statistical testing with proper data visualization can easily be streamlined. These combinations have considerable potential to improve the accuracy and statistical power [11] of the analysis of RNA-Seq gene expression data.

Results

Fig. 1 shows the five steps of TRAPR, i.e., data manipulation, data preprocessing, statistical analysis, preprocessing result visualization, and statistical result visualization.

Data manipulation

TRAPR provides two functions to import RNA-Seq experimental data and four functions to export results to files. TRAPR can read text files for expression data as well as for a list of genes. During or following analysis, users can export DEG lists or detailed tables for DEG and expression tables, which other tools can utilize.

Data preprocessing

TRAPR provides three types of data preprocessing methods: filtering, transformation, and normalization. TRAPR filtering has six filter types: sample, gene, zero value, low variance, low expression, and gene list. Unlike DNA microarrays that have a fixed number of probes, RNA-Seq explores massive amounts of isoforms and novel transcripts mixed with noise, such that it returns many zeros and nonsense values. Genes encoding miRNAs or snoRNAs often show extremely high expression levels, even though they are treated by a poly-A purification procedure. These outliers can easily be removed by zero-value and gene filters. Statistical power can be improved by low-expression and low-variance filters by reducing non-standard distributions. Analyzing different combinations of samples can conveniently be supported by sample filtering.

TRAPR provides two well-known transformation functions, log2 transformation and VSN, followed by hyperbolic arcsin, arcsin(x), and transformation [11] to standardize data distribution and normalize variance distribution, respectively.

TRAPR provides many normalization methods, including upper quantile [12], quantile, mean, and median normalizations. One can conveniently compare the effect of applying

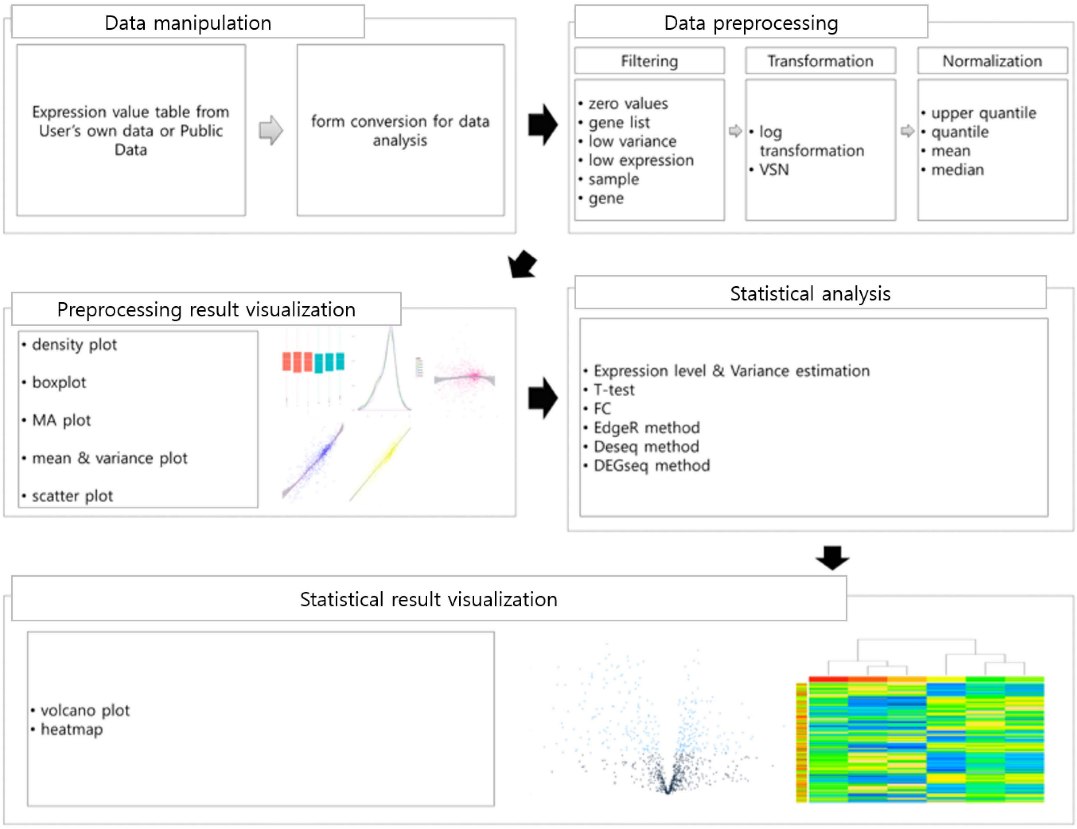

Fig. 1. Five steps of typical RNA sequencing analysis pipeline, which are included in the TRAPR package. VSN, variance stabilizing normalization.

different methods with the following statistical testing and visualization functions.

Statistical testing

TRAPR has several statistical testing functions to identify DEGs. Student t-test and statistical methods suggested in edgeR, baySeq, and DESeq assume a normal distribution or a Poisson distribution. Meanwhile, methods in DEGseq and NOISeq [12] do not need to make any assumption. For now, there is no correct answer that fits the characteristics of RNA-Seq data, and each method has its own merits. TRAPR allows users to choose their own methods to build customized analysis pipelines. The t test is recommended for large datasets, whereas edgeR and simple-fold change work for datasets with a small number of samples. DEGs can be labeled and saved as files containing lists of gene names with detailed information.

Visualization

Data preprocessing steps are not supported by visualization functions in previously developed packages, while proper visualization is essential and powerful for evaluating the quality of the RNA-Seq data and the preprocessing steps. TRAPR provides five flexible plotting functions, including density, boxplot, MA, scatter, and mean–variance plots. Volcano plots and heatmaps are also provided to visualize the results of statistical analysis. Each visualization function has direct access to FPKM values and differential expression values.

Discussion

We have developed TRAPR, an R package for RNA-Seq data analysis. TRAPR provides an entire pipeline for RNA-Seq analysis, which is not merely a combination of currently available tools, but the backbone that facilitates the proper application and coordination of these tools. For instance, upper-quartile normalization followed by zero-value filtering, VSN, and edgeR statistical testing with proper data visualization can easily be streamlined through TRAPR. These combinations will help improve accuracy and statistical power. TRAPR provides visualization tools and file I/O functions to evaluate the quality and characteristics of the data. TRAPR was developed and integrated in R, such that it can be easily applied to other technologies like Serial Analysis of Gene Expression and microarray. Various filters have been integrated into the package. TRAPR can be used as a platform to interweave RNA-Seq data analysis tools and packages to take advantage of the virtues of each.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2012-0000994), the Korean Health Technology R&D Project, the Ministry of Health and Welfare (HI13C2164), and the ICT R&D program of MSIP/IITP [B0101-15-247 for the “Development of Open ICT Healing Platform Using Personal Health Data”], a grant (16183-MFDS541) from Ministry of Food and Drug Safety in 2016.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57-63.
2. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* 2011;8:469-477.
3. Auer PL, Srivastava S, Doerge RW. Differential expression: the next generation and beyond. *Brief Funct Genomics* 2012; 11:57-62.
4. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; 5:R80.
5. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139-140.
6. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11:R106.
7. 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;26:136-138.
8. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010;11:422.
9. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. *Nucleic Acids Res* 2007;35:e102.
10. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002;18 Suppl 1:S96-S104.
11. Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. *Algorithms Mol Biol* 2012;7:5.
12. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011;21:2213-2223.