Bioinformatics

doi.10.1093/bioinformatics/xxxxxx

Advance Access Publication Date: Day Month Year

Manuscript Category



Subject Section

Visualization methods for RNA-sequencing data analysis

Lindsay Rutter 1,*, Michelle Graham 2 and Dianne Cook 3,*

- ¹ Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA,
- ²Department of Agronomy, Iowa State University, Ames, IA 50011, USA and
- ³Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: It was initially claimed that RNA-seq produced unbiased data that did not require sophisticated normalization. However, studies have since revealed that RNA-seq data is biased and that accurate detection of differentially expressed genes is not a trivial task. In light of these findings, researchers should analyze RNA-seq data like they would any other biased multivariate data. The most effective approach to modern data analysis is to iterate between models and visuals, and to enhance the appropriateness of models based on feedback from visuals. Unfortunately, researchers do not often use models and visuals in a complimentary fashion when analyzing RNA-seq data.

Results: We use real RNA-seq data to show that our visualization tools can detect normalization problems, DEG designation problems, and common errors in RNA-seq analysis. We also show that our tools can identify genes of interest that cannot be obtained by models. In this paper, we do not propose that users radically change their approach to RNA-seq analysis. Instead, we propose that users simply modify their approach to RNA-seq analysis by assessing the sensibility of their models with multivariate statistical graphics.

Contact: lrutter@iastate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-sequencing (RNA-seq) uses next-generation sequencing (NGS) to estimate the quantity of RNA in biological samples at given timepoints. In recent years, decreasing cost and increasing throughput has rendered RNA-seq an attractive alternative to transcriptome profiling. Prior to RNA-seq, gene expression studies were performed with microarray techniques, which required prior knowledge of reference sequences. RNA-seq does not have this limitation, and has enabled a new range of applications such as transcriptome de novo assembly (Robertson *et al.*, 2010) and detection of alternative splicing processes (Anders *et al.*, 2012; Pan *et al.*, 2008). Coupled with its high resolution and sensitivity, RNA-seq will likely revolutionize our understanding of the intricacies of eukaryotic transcriptomes (Wang *et al.*, 2009; Zhao *et al.*, 2014).

RNA-seq data is multivariate data, and its basic form is a matrix containing mapped read counts for *n* rows of genes and *p* columns of samples. These mapped read counts provide estimations of the gene expression levels across samples. Researchers typically conduct RNA-seq studies to identify differentially expressed genes (DEGs) between treatment groups. In most popular RNA-seq analysis packages, this objective is approached with models, such as the negative binomial model (Anders and Huber, 2010; Trapnell et al., 2013; Trapnell et al., 2012; Robinson et al., 2010) and linear regression models (Law et al., 2014).

Initially, it was widely claimed that RNA-seq produced unbiased data that did not require sophisticated normalization (Wang et al., 2009; Morin et al., 2008; Marioni et al., 2008). However, numerous studies have since revealed that RNA-seq data is replete with biases and that accurate detection of DEGs is not a negligible task. Problems that complicate the analysis of RNA-seq data include nucleotide and read-position biases

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

^{*}To whom correspondence should be addressed.

2 Sample et al.

(Hansen et al., 2010), biases related to gene lengths and sequencing depths (Oshlack, Robinson, and Young, 2010; Robinson and Oshlack, 2010), biases introduced during library preparation (McIntyre et al., 2011), biases pertaining to the number of replications (Schurch et al., 2016), biases derived from overlapping sense-antisense transcripts and gene isoforms (Trapnell et al., 2013), and the confounding combination of technical and biological variability (Bullard et al., 2010).

In light of these complications, researchers should analyze RNA-seq data like they would any other biased multivariate data. Simply applying models to such data is problematic because models hold assumptions that they alone cannot call into question. Fortunately, data visualization enables researchers to see patterns and problems they may not otherwise detect with traditional modeling. As a result, the most effective approach to data analysis is to iterate between models and visuals, and enhance the appropriateness of applied models based on feedback from visuals (Shneiderman, 2002). With RNA-seq data, we primarily want to compare the variability between replicates and between treatment groups. This is visually best achieved by drawing the mapped read count distributions across all genes and samples. Unfortunately, the few plotting tools offered in popular RNA-seq packages do not allow users to effectively view their data in this manner.

In this paper, we strive to remedy this problem by publishing new and effective RNA-seq plotting tools. We use real RNA-seq data to show that our tools can detect normalization problems, DEG designation problems, and common errors in the analysis pipeline. We also show that our tools can identify genes of interest that cannot otherwise be obtained by models. We emphasize that interactive graphics should be an indisposable component of modern RNA-seq analysis: Researchers should be able to quickly flip through plots of genes that appear promising or problematic, and link between plots to swiftly obtain various perspectives of their data. Here, we do not propose that users drastically change their approach to RNA-seq analysis. Instead, we propose that users simply modify their approach to RNA-seq analysis by assessing the sensibility of their models with multivariate graphical tools, namely with parallel coordinate plots, scatterplot matrices, and replicate point plots.

2 Approach

3 Methods

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text. ? might want to know about text text text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. ? might want to know about Text.

3.1 This is subheading

Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text. ? might want to know about text text text text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. ? might want to know about Text. Figure 2 shows that the above method Text Text. ? might want to know about text text text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text. ? might want to know about text text text Text Text

3.1.1 This is subsubheading

Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text. ? might want to know about text text text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text. Figure 2 shows that the above method Text Text. ? might want to know about text text text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text. ? might want to know about text text text Text Text Text Text Text Text.

4 Sample et al.

Table 1. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

FPO

Fig. 1. Caption, caption

3.2 Test1

4 Discussion

5 Conclusion

- 1. this is item, use enumerate
- 2. this is item, use enumerate
- 3. this is item, use enumerate

Acknowledgements

Text Text Text Text Text Text Text Text. ? might want to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data, Genome Biology, 11, R106.
- Anders, S., Reyes, A., and Huber, W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22, 2008–2017.
- Baggerly, K.A. and Coombes, K.R. (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, **3**, 1309–1334.
- Brown, A.V. and Hudson, K.A. (2015) Developmental profiling of gene expression in soybean trifoliate leaves and cotyledons. BMC Plant Biology, 15, 169.
- Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 94.
- Cook, D., Hofmann, H., Lee, E.-K., Yang, H., Nikolau, B., and Wurtele, E. (2007) Exploring gene expression data, using plots. *Journal of Data Science*, 5, 151–182.
- Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38, e131.
- Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., and Carvalho, B.S. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12, 115–121.
- Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G.P., Petretto, E. and van Noort, V. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41, 149–155.
- Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15, R29.
- Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008)
- 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*, 18, 1509–1517.
- McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J., et al. (2011) RNAseq: Technical variability and sampling. *BMC Genomics*, **12**, 293.
- Moran Lauter, A.N., Peiffer, G.A., Yin, T., Whitham, S.A., Cook, D., and Shoemaker, R.C. (2014) Identification of candidate genes involved in early iron deficiency chlorosis signaling in soybean (glycine max) roots and leaves. *BMC Genomics*, 15, 1–25.
- Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., et al. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, **45**, 81–94.
- Oshlack, A., Robinson, M.D., and Young, M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biology*, 11, 220.
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40, 1413–1415.
- Risso, D., Schwartz, K., Sherlock, G., Dudoit, S. (2011) GC-Content normalization for RNA-Seq data. BMC Bioinformatics, 12, 480.

6 Sample et al.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.
Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et

- al. (2010) De novo assembly and analysis of RNA-seq data. Nature Methods, 7, 909-912.
- Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11, R25.
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, **26**, 139–140.
- Shneiderman, B. (2002) Inventing discovery tools: Combining information visualization with data mining. *Information Visualization*, 1, 5–12.
 Schurch, N.J., Schofield, P., GierliÅ, ski, M., Cole, C., Sherstnev, A., Singh, V., et
- al. (2016) How many biological replicates are needed in an RNA-seq experiment

- and which differential expression tool should you use? RNA, 22, 839-851.
- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. C., Cornell M., C., Cornell M., Cornel(2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**, 46–53.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., and Kelley D.R. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, **7**, 562–578.
- Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014) Comparison
- of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE, 9, e78644.