

## Visualization methods for RNA-sequencing data

SUMMARY: This is the summary for this paper.

KEY WORDS: Data visualization; Exploratory data analysis; Interactive graphics; RNA-sequencing; Statistical graphics

## 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 (Grabherr et al., 2011; Robertson et al., 2010) and detection of alternative splicing processes (Anders, Reyes, and Huber, 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, Gerstein, and Snyder, 2009; Zhao et al., 2014).

Gene expression 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-specific and read-position specific biases (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 senseantisense 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 have 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 analyze data is to iterate between models and visuals, and to enhance the appropriateness of applied models based on feedback from visuals (Unwin, 1992; Shneiderman, 2002).

When visualizing RNA-seq data, we primarily want to compare the variability between replicates and between treatment groups. This is best achieved by viewing the mapped read count distributions across all genes and samples. Unfortunately, the few visual tools available in popular RNA-seq packages do not allow users to effectively view their data in this manner.

Here, we use real RNA-seq data to demonstrate that our visualization 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 any models. In line with modern multivariate data exploration, we emphasize that interactive graphics should be an indispensable component of RNA-seq analysis. Researchers should be able to quickly flip through plots of genes that appear promising or problematic, and link between different types of plots to quickly obtain various perspectives of their data. In this paper, we are not proposing that users completely 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 parallel coordinate plots, scatterplot matrices, and replicate line plots.

## 2. Parallel Coordinate Plots

There are several packages within the BioConductor software that provide graphics for RNA-seq data analysis (Huber et al., 2015). Two of the most common graphic techniques are the side-by-side boxplot and Multidimensional scaling (MDS) plot (Love, Huber, and Anders, 2014; Risso et al., 2011; Robinson et al., 2010; Su et al., 2016; Ritchie et al., 2015; Marini, 2017). Unfortunately, these plots can hide problems that still exist in the data even after normalization, and that could be better detected with parallel coordinate plots.

The image above exemplifies this problem for a dataset containing five replicate samples, each with 50 random values drawn from a normal distribution centered at a value of 10. The 50 values from each replicate were sorted from smallest to largest. However, for the images in the left half of the plot, the second and fourth replicate values were sorted from largest to smallest. We cannot see this alteration in the data when we compare the left and right boxplots. This is because the boxplots do not show sample connections between each of the 50 observations. We can see that there may have been an alteration in the data when we compare the left and right MDS plots. Namely, the left MDS plot contains a clustering structure that suggests the second and fourth replicates are similar compared to the first, third, and fifth replicates. However, we still cannot ascertain why these replicates are different.

We can see the most information about the alteration in the data when we compare the left and right parallel coordinate plots - because we can now judge how consistent the replicate values are for each of the 50 observations. We notice that certain observations (the pinks and oranges) are inconsistent (crossing) between replicates in the left plot, while all observations are consistent (level) between replicates in the right plot. Analysts should ensure (rather

than assume) that replicate values are as consistent as intended, and they can verify this most clearly with the parallel coordinate plots.

[Figure 1 about here.]

You can experiment with fancier figures than Figure 1.

[Table 1 about here.]

You can experiment with fancier tables than Table 1.

We can get bold symbols using `\bmath`, for example,  $\alpha_i$ .

### 3. Discussion

Put your final comments here.

#### ACKNOWLEDGEMENTS

The authors thank Professor A. Sen for some helpful suggestions, Dr C. R. Rangarajan for a critical reading of the original version of the paper, and an anonymous referee for very useful comments that improved the presentation of the paper.

#### SUPPLEMENTARY MATERIALS

#### 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**, 20082017.
- 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.

- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652.
- 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.
- 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.
- Marini, F. (2017). pcaExplorer: Interactive Visualization of RNA-seq Data Using a Principal Components Approach. GitHub, Inc. <https://github.com/federicomarini/pcaExplorer> (accessed October 7, 2017).
- 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.
- 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**, 819-824.
- 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**, 14131415.
- Risso, D., Schwartz, K., Sherlock, G., Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12**, 480.
- 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**, 909912.
- 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**, 139140.
- Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization* **1**, 5-12.
- Schurch, N.J., Schofield, P., Gierliski, 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**, 839851.
- Shu, S., Ritchie, M.E., Law, C., and Lee, S. (2016). Glimma: Interactive HTML graphics. GitHub, Inc. <https://github.com/Shians/Glimma/> (accessed October 7, 2017).
- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**, 4653.

- 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**, 562578.
- Unwin, A. (1992). How interactive graphics will revolutionize statistical practice. *The Statistician* **41**, 365-369
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 5763.
- 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.

## APPENDIX

*Title of appendix*

Put your short appendix here. Remember, longer appendices are possible when presented as Supplementary Web Material. Please review and follow the journal policy for this material, available under Instructions for Authors at <http://www.biometrics.tibs.org>.



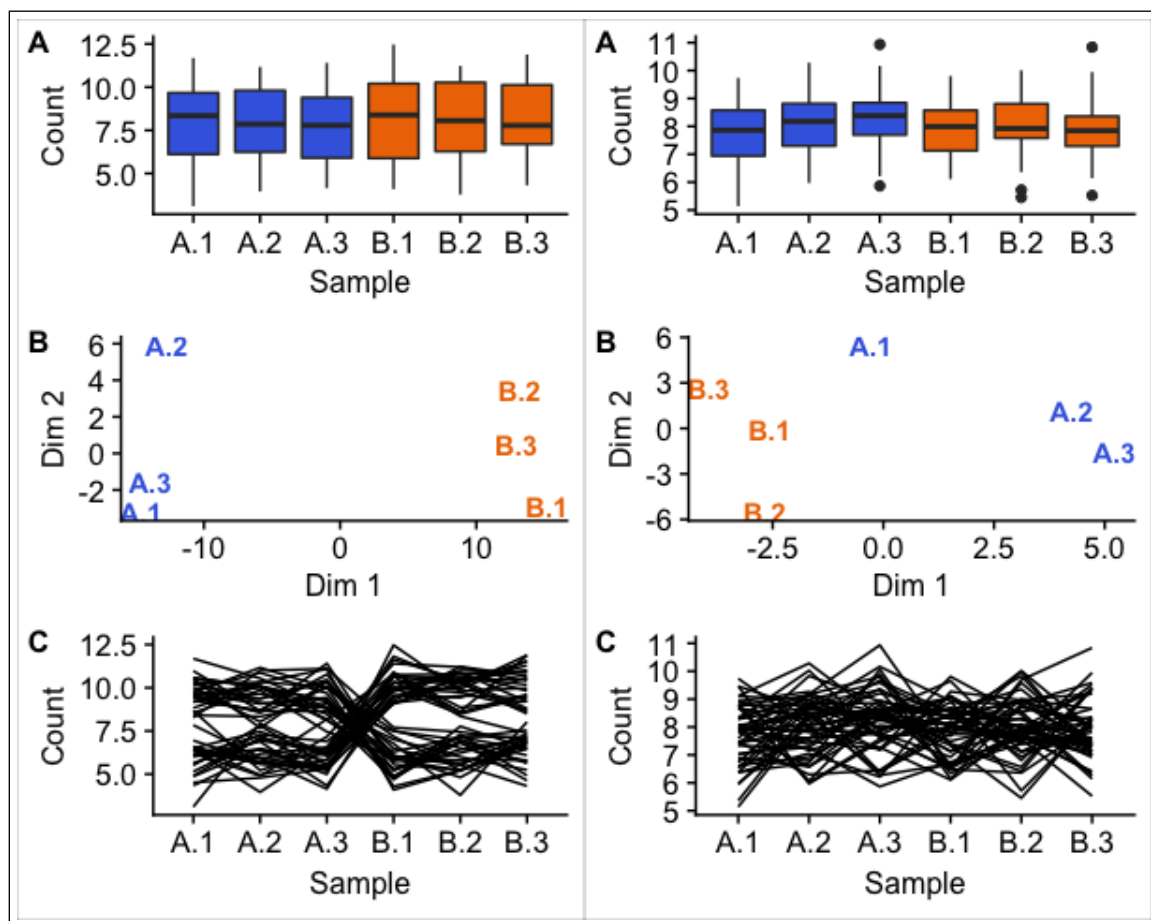


Figure 1. Caption.

**Table 1***This is a simple table.*

Culture/Library prep.	Library prep. protocol	Growth condition	Flow-cell
Y1	Protocol 1	YPD	428R1
Y1	Protocol 1	YPD	4328B
Y2	Protocol 1	YPD	428R1
Y2	Protocol 1	YPD	4328B
Y7	Protocol 1	YPD	428R1
Y7	Protocol 1	YPD	4328B
Y4	Protocol 2	YPD	61MKN
Y4	Protocol 2	YPD	61MKN