Shiny-Seq: advanced guided transcriptome analysis

Zenitha Sundararajan
Hochschule Bonn-Rhein-Sieg
Rainer Knoll
Hochschule Bonn-Rhein-Sieg
Peter Hombach
Hochschule Bonn-Rhein-Sieg
Matthias Becker
Deutsches Zentrum fur Neurodegenerative Erkrankungen
Joachim L. Schultze
Hochschule Bonn-Rhein-Sieg
Thomas Ulas (t.ulas@uni-bonn.de)
LIMES https://orcid.org/0000-0002-9785-4197

Research note

Keywords: RNA-Seq, Bioinformatics, Analysis, Shiny, DeSeq2, functional prediction, Limma, co-expression network analysis, pipeline, automated report

Posted Date: June 27th, 2019

DOI: https://doi.org/10.21203/rs.2.10701/v1

Version of Record: A version of this preprint was published on July 18th, 2019. See the published version at https://doi.org/10.1186/s13104-019-4471-1.

Abstract

Objective Over the last years, Next Generation Sequencing has generated huge amounts of gene expression data supporting research to answer biological and medical questions. While several tools and advanced guided analysis pipelines were designed for this purpose, some require programming skills and others lack the support for many important features that enable more comprehensive data analysis. Results We present the tool Shiny-Seq. It has many new features such as batch effect evaluation and removal, quality check with several visualization options, enrichment analysis with multiple biological databases, identification of patterns using advanced methods such as weighted gene co-expression network analysis, summarizing analysis as power point presentation and all results as tables via one click feature. The source code is published on GitHub (https://github.com/schultzelab/Shiny-Seq) with the code licensed under GPLv3. Shiny-Seq is written in R using the Shiny framework. In addition, the application is hosted on a public website hosted by shinyapps.io server (https://schultzelab.shinyapps.io/Shiny-Seq/) and as a Docker image https://hub.docker.com/r/makaho/shiny-seq.

Introduction

The scientific community is continuously trying to improve their understanding of genetic mechanisms in biological systems in a global way. Particularly transcriptome analysis is a major work horse to assess regulation and function of complete genomes [1]. Here, Next Generation Sequencing (NGS) has become one of the preferred methods. Constantly dropping sequencing costs and more than 25000 (ArrayExpress, NCBI GEO) publically available transcriptome datasets help us to better understand the complex relationship between genotype and phenotype. With growing accessibility, still, only the minority of investigators in the life and medical sciences has the means to analyze and leverage this enormous treasure of data. Understanding RNA-Seq data requires several successive steps in order to analyze, visualize and interpret it. The key steps are (i) import of data, (ii) normalization, (iii) analysis using statistical techniques such as hypothesis testing, (iv) functional enrichment analysis using various biological databases, and (v) identification of biological patterns using advanced methods such co-expression network analysis. Integrated, simply accessible, easily expandable and inexpensive tools are still missing. Shiny-Seq is providing such an analysis environment for the broader community in the life and medical sciences.

Main Text

Shiny-Seq features

In the following, we provide details regarding features implemented in the various steps of Shiny-Seq. The main text consists of three different sections: data pre-processing (1), exploratory data analysis (2), and downstream analysis (3) and its respective subsections.

Data pre-processing (1)

Input (1.1)

Our Shiny-Seq pipeline provides two different starting points for the analysis. First, the count table, which is the universal file format produced by most of the alignment and quantification tools. Second, the transcript-level abundance estimates provided by ultrafast pseudoalignment tools like *Kallisto* [2]. For this purpose, the user has to provide the location of the directory containing the files generated by *Kallisto*. Another essential input is the annotation file, a matrix that stores different factors associated with each sample.

Normalization (1.2)

The package *DESeq2* [3] normalizes the dataset by computing a size factor for each sample. The size factor is calculated by taking the median ratio of each sample over a reference or pseudo sample. Shiny-Seq uses the default parameter recommended by the Bioconductor DESeq2 workflow for RNA-Seq [4] data but also allows to control for $\log_2$ fold change shrinkage and multiple testing, custom p-value and fold change cut-offs.

Batch effect analysis (1.3)

Batch effects can be induced by either known variables such as technical heterogeneity and time of experiment or by unknown variables [5]. In Shiny-Seq, we use *removeBatcheffect* from *LIMMA* [6] to account for the batch effect from known sources. For unknown variables we use *SVA* [5]to construct surrogate variables to account for technical variability. The influence of potential variables known to cause the batch effect can then be examined by PCA. The detected batch effects are modeled within the *DESeq2* study design and the batch corrected data is used for all respective visualizations.

Additionally, Shiny-Seq can estimate the influence of the batch effect based on an ANOVA model and visualize it via a source of variation plot showing the effects sizes of the modeled factors.

Exploratory data analysis (2)

Differential gene expression analysis (2.1)

Shiny-Seq supports *DeSeq2*'s differential gene expression testing (DGEA) based on a negative binomial distribution model. *DeSeq2* uses variance-mean estimation for RNA-Seq data and the Wald test. The Wald test assumes that the Z-statistic takes a standard normal distribution with zero mean and unit variance. Additionally, Shiny-Seq supports p-value evaluation and correction, where a histogram is generated, which

helps to decide whether the statistical hypothesis assumption is violated. If necessary the correction can be performed using *fdrtool* [7].

Co-expression network analysis (2.2)

In contrast to conventional DGEA Shiny-Seq also provides a co-expression network analysis (CENA) function using WGCNA [8]. This method allows identifying modules based on correlation followed by network analysis. It takes the pre-processed data and the annotation file as inputs but can also take results from the DEGA as starting point. Note that batch corrected data is used as input for the CENA if a batch correction was selected beforehand. The output is the typical module-condition relationship heat map and a table including module name, number of genes and identified hub genes in each module. Furthermore, the identified modules are integrated into Shiny-Seq in a way that the user can perform most of the downstream analysis e.g. functional enrichment analysis, heat maps, and Venn diagrams based on these results.

Downstream analysis (3)

Functional prediction (3.1)

After DGEA and CENA a functional prediction based on gene set enrichment analysis (GSEA) can be performed. Shiny-Seq uses biological databases such as KEGG [9], GO [10] and Broad's molecular signatures database (MSigDB) [11] in *clusterprofiler*'s [12] GSEA to take advantage of already publicly available knowledge, which assists during the interpretation process.

Transcription factor binding side overrepresentation analysis (3.2)

Our application also performs a transcription factor binding site overrepresentation analysis in the promoter regions for all groups of genes being identified by DGEA and CENA. Predicted transcription factors are marked in the table of differentially expressed genes. This analysis provides valuable information about potential upstream regulators responsible for the observed genotype. Shiny-Seq uses *pcaGopromoter* [13] to predict transcription factors.

Visualization (3.3)

Shiny-Seq provides a multitude of visualizations in the respective analysis steps (Supp. Fig. 1). This include plots such as heat maps and volcano plots, which are commonly used during the analysis of RNA-Seq data. A heat map for example visualizes relationships between samples and genes. In Shiny-Seq we use heat maps for the visualization of differentially expressed genes, 1000 genes having the highest variance within the data and all present and differential expressed transcription factors. Volcano

plots help to visualize differentially expressed genes obtained from DGEA. While heat maps and volcano plots are used to visualize e.g. hypothesis test results of a single comparison, they do not have the capability to compare results obtained from multiple comparisons. Shiny-Seq addresses this through a Venn diagram and a fold change fold change plot, where the names of genes of interest can be identified by selecting them interactively in the respective plot. Static plots e.g. heat maps can be download as vector graphic for further usage. If meaningful, some of the plots can be further customized within Shiny-Seq.

Figure 1: Data pre-processing (A): Box plots of samples (before and after normalization), PCA (2D and 3D) of samples (before, after normalization and after batch correction; interactive), sample correlation plot (before and after batch correction), source of variation plot (before and after batch correction; interactive); Exploratory analysis (B): box plot of single gene expression including statistics, p-value evaluation histogram, MA plot, module-condition relationship heat map (CENA), Venn diagram (interactive), volcano plot (interactive), fold change fold change plot (interactive), heatmap of 1000 most variable genes, own gene list, DEGA and CENA results; Downstream analysis (C): dot plots of GSEA results (interactive), visualization of KEGG pathways (DEGA genes or all present genes), TFBS plot.

Generation of report (3.4)

Another unique feature is the compilation of all outputs generated during each step of the analysis and summarizing these results in a PowerPoint presentation, as well as respective tables, which can be downloaded and shared with colleagues and collaborators. It includes QC plots e.g. box plots and PCA plots before and after normalization, top-10 up-regulated and down-regulated genes, enrichment analysis results. The R package *ReporteRs* [14] is used to generate a presentation.

Conclusion

Global transcriptome analysis has become a standard approach in research but also in clinical settings. At the same time, experts who can analyse this kind of data are still the limiting factor. Shiny-Seq provides a framework for analysing such data in a transparent and reproducible manner for NGS service providers, NGS competence centres, but also for end users with limited scripting experience. It offers a huge functionality combined with a guided and intuitive workflow and a comprehensive and time saving summary functionality.

Limitations

While the development is complete from the end-user perspective, the internally used R codes are still cluttered. Moreover, incorporation of new features and further customization of the visualizations would furthermore improve Shiny-Seq. The application currently supports only enrichment analysis of gene ontologies, pathways, and molecular signatures. We intend to extend support to disease ontologies as well. We also intend to support preparation of count table from transcript quantification files generated by

other tools such as Star [15], HTSeq-counts [16], and Sailfish [17]. The export of a DESeq2 rData object would provide more flexibility for users with programming experience.

Abbreviations

DGEA – Differentially genes expression analysis

CENA – Co-expression network analysis

GO – Gene ontology

GSEA - Gene set enrichment analysis

KEGG - Kyoto Encyclopedia of Genes and Genomes

PCA – Principal component analysis

QC – Quality control

MSigDB - Molecular signatures database

NGS - Next Generation Sequencing

WGCNA – Weighted gene co-expression network analysis

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data in this study were included in this article.

Competing interests

The authors declare that they have no competing interests.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC2151/1 – 390873048. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

ZS and RK developed the Shiny-Seq software, drafted the manuscript and provided critical comments as the first authors. PH and MB debugged and helped to improve the tool and provided critical comments to the paper. JLS supervised the critical discussion and revised the paper. TU made the conception of the software, substantially revised the paper, and led this project to completion. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the group of Prof. Dr. Joachim L. Schultze and the LIMES institute for testing and reviewing of Shiny-Seq.

Publisher's Note

Open Access

References

[1] N. Shirley, "Transcriptomics technologies," *PLOS Comput. Biol.*, vol. 13, no. 5, p. e1005457, 2017.

[2] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification.," *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–7, 2016.

[3] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, 2014.

[4] M. I. Love, S. Anders, and W. Huber, “Analyzing RNA-seq data with DESeq2.” 2019.

[5] J. T. Leek and J. D. Storey, “Capturing heterogeneity in gene expression studies by surrogate variable analysis,” *PLoS Genet.*, 2007.

[6] M. E. Ritchie *et al.*, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, 2015.

[7] K. Strimmer, “fdrtool: A versatile R package for estimating local and tail area-based false discovery rates,” *Bioinformatics*, vol. 24, no. 12, pp. 1461–1462, 2008.

[8] P. Langfelder and S. Horvath, “WGCNA: An R package for weighted correlation network analysis,” *BMC Bioinformatics*, 2008.

[9] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*. 1999.

[10] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*. 2000.

[11] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[12] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, “clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters,” *Omi. A J. Integr. Biol.*, 2012.

[13] T. A. Gerds, O. H. Nielsen, M. Hansen, J. Olsen, J. T. Troelsen, and J. B. Seidelin, “pcaGoPromoter - An R Package for Biological and Regulatory Interpretation of Principal Components in Genome-Wide Gene Expression Data,” *PLoS One*, 2012.

[14] T. D. Gohel, “ReporteRs package manual to generate PowerPoint presentation.” 2017.

[15] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, 2013.

[16] S. Anders, P. T. Pyl, and W. Huber, “HTSeq-A Python framework to work with high-throughput sequencing data,” *Bioinformatics*, 2015.

[17] R. Patro, S. M. Mount, and C. Kingsford, “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms,” *Nat. Biotechnol.*, 2014.

Figure Legend

Figure 1: Data pre-processing (A): Box plots of samples (before and after normalization), PCA (2D and 3D) of samples (before, after normalization and after batch correction; interactive), sample correlation plot (before and after batch correction), source of variation plot (before and after batch correction; interactive); Exploratory analysis (B): box plot of single gene expression including statistics, p-value evaluation histogram, MA plot, module-condition relationship heat map (CENA), Venn diagram (interactive), volcano plot (interactive), fold change fold change plot (interactive), heatmap of 1000 most variable genes, own gene list, DEGA and CENA results; Downstream analysis (C): dot plots of GSEA results (interactive), visualization of KEGG pathways (DEGA genes or all present genes), TFBS plot.

Figures

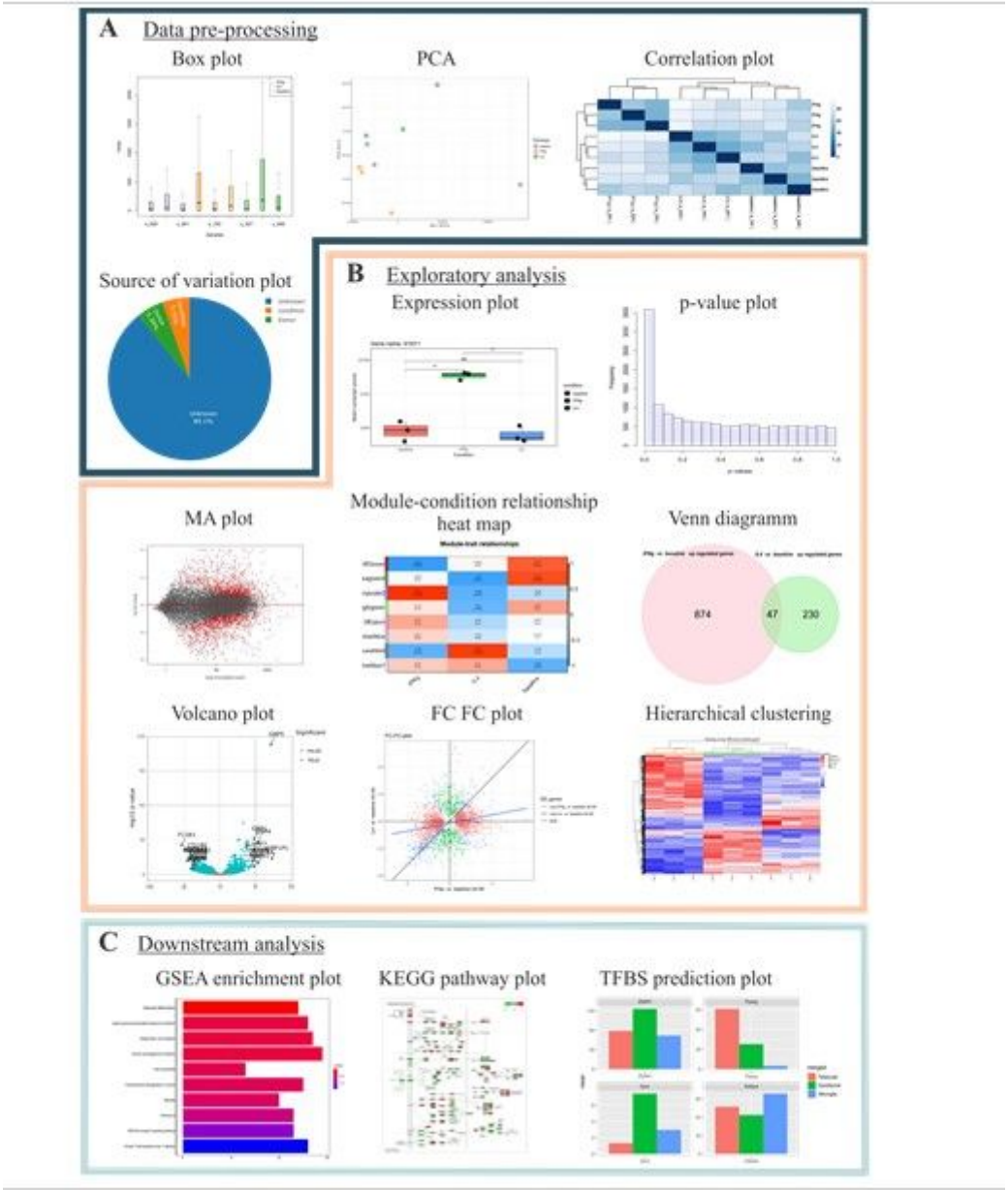

Figure 1

Data pre-processing (A): Box plots of samples (before andafter normalization), PCA (2D and 3D) of samples (before, after normalization and after batch correction; interactive), sample correlation plot (before and after batch correction), source of variation plot (before and after batch correction; interactive);Exploratory analysis (B): box plot of single gene expression including statistics, p-value evaluation histogram, MA plot, module-condition relationship heat map (CENA), Venn diagram (interactive), volcano plot (interactive), fold change fold change plot (interactive), heatmap of 1000 most variable genes, own gene list, DEGA and CENA results; Downstream analysis (C): dot plots of GSEA results (interactive), visualization of KEGG pathways (DEGA genes or all present genes), TFBS plot.