freeCount: A Coding Free Framework for Guided Count Data Visualization and Analysis

Elizabeth M. Brooks
University of Notre Dame
Notre Dame, Indiana, USA
ebrooks5@nd.edu

Sheri A. Sanders
University of Notre Dame
Notre Dame, Indiana, USA
ssander5@nd.edu

Michael E. Pfrender
University of Notre Dame
Notre Dame, Indiana, USA
mpfrende@nd.edu

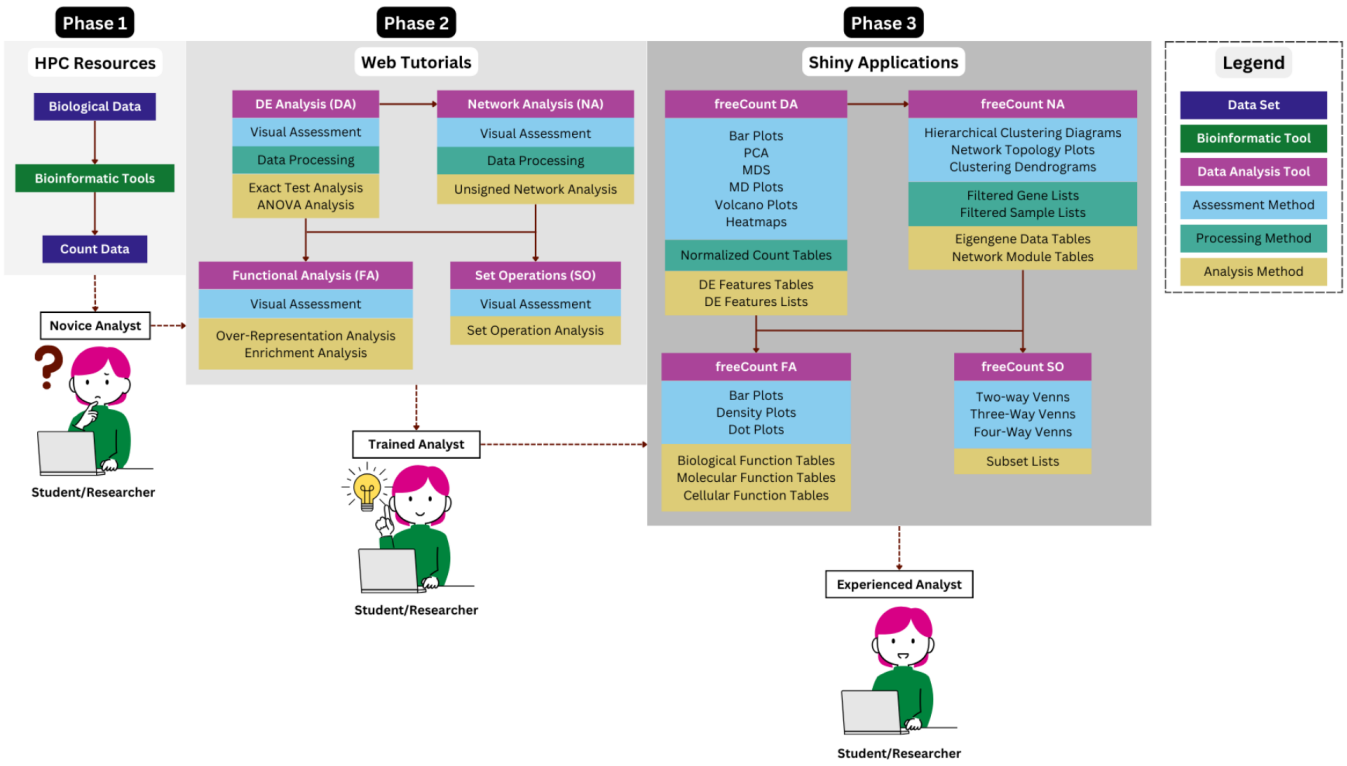

Figure 1: Application user journey diagram. The freeCount user journey diagram depicts the user-centered process of count data analysis. The typical user is a student or researcher who begins their journey as a novice analyst. In the first phase the novice analyst has received count data, but they are unsure how to proceed. The second phase is where the analyst completes our web tutorials and now has an idea about how to proceed with analyzing their data. In the third phase the analyst uses our Shiny applications to produce high-quality analysis results.

ABSTRACT

The analysis and interpretation of high-dimensional biological data sets is a challenging task. Exploratory data analysis of count data produced by next-generation sequencing technologies presents a common hurdle to researchers. Biologists often find it difficult to get started with the analysis process, which can be time consuming and repetitive. With the freeCount analysis framework students and researchers are guided through the iterative steps of data assessment, processing, and analysis in a visual environment. The freeCount analysis framework takes advantage of the reactive features of R Shiny to deliver a set of modular and interactive tools and tutorials for the structured analysis and visualization of count data.

PEARC '24, July 21–25, 2024, Providence, RI, USA

ACM ISBN 979-8-4007-0419-2/24/07
https://doi.org/10.1145/3626203.3670605

CCS CONCEPTS

• **Human-centered computing** → **Visualization systems and tools**; *User centered design*; • **Applied computing** → **Education**; **Genetics**; **Bioinformatics**; **Genomics**; **Computational biology**.

KEYWORDS

R Shiny, Interactive Interfaces, Analysis Framework, Genomics, Transcriptomics, Metagenomics

ACM Reference Format:

Elizabeth M. Brooks, Sheri A. Sanders, and Michael E. Pfrender. 2024. freeCount: A Coding Free Framework for Guided Count Data Visualization and Analysis. In *Practice and Experience in Advanced Research Computing (PEARC '24), July 21–25, 2024, Providence, RI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626203.3670605

1 INTRODUCTION

The freeCount analysis framework provides a modular set of tools and tutorials to guide users through a structured approach to biological data analysis (https://github.com/ElizabethBrooks/freeCount). In this paper we outline the utility and features of the freeCount analysis framework, which has been used by researchers and students at the University of Notre Dame (ND). The purpose of our project was to provide students and researchers with a coding-free framework for the rapid analysis of count data. Count data is produced by processing a variety of routinely generated omics data (e.g., transcriptomics, genomics, metagenomics). While the methods were originally designed for analyzing gene expression patterns measured by bulk RNA sequencing (RNAseq), the methods have been applied more broadly to single cell RNA sequencing (scRNAseq) and metagenomics data. Count data tables consist of rows that represent a set of features (e.g. genes, genome features, or species) and columns with the recorded features (counts) for each sample.

Data assessment and processing is an important first step of any analysis. The goal is to remove sources of technical bias (e.g., library size, missing data) through normalization and filtering. Visualization of the data is frequently necessary to facilitate the removal of major outliers. Additionally, data visualization methods enable researchers to assess the need for more data. Frequent approaches to analyzing count data include differential expression (DE), network, and functional analysis. In DE analysis samples are grouped by condition (e.g., treatment, cell type, sample), then averages and distributions are calculated. The resulting patterns are used to assess the statistical likelihood of differences between groups. In the analysis of RNA or scRNA transcriptomic data, DE analysis allows researchers to identify genes that are differentially expressed. In the case of metagenmoics data, DE analysis can be used to determine the bacteria species that are more or less numerous. The results of DE analysis are used to identify pathways or communities that are impacted by the study conditions, finally getting back to the biological question.

The exploratory data analysis process is repetitive and iterative, requiring extensive amounts of code and small adjustments to thresholds and filters throughout. Navigating the required amount of code is a frequent barrier for users that contributes to significant delays in the evaluation and interpretation of count data. The freeCount analysis framework utilizes R Shiny [3], bioconductor [9], and base R [15] packages to create interactive tools with graphical user-interfaces. R and bioconductor were chosen as they are the most common platform and tools currently used by the community for the analysis and visualization of these biological data sets. Our user-friendly applications are designed to be easy to run, easy to learn, and easy to scale.

2 FEATURES

2.1 Running & Scaling

The freeCount applications can be run locally in R or Rstudio, as a single Shiny server, or as a swarm of virtual machines (VMs). freeCount can be installed from github, with dependency handling built in, and then launched as any other R Shiny application. Detailed installation instructions are available on the GitHub website, along with tutorials for running each of the applications. This instance provides rapid, interactive data analysis that reduces hundreds of lines of code to a handful of commands to launch.

A Shiny server version of freeCount is available on the Jetstream2 [7] image library, allowing for ACCESS users to launch an instance and run freeCount through a browser, and with code fully eliminated from data exploration. A second image is also available on the image library, which quickly sets up a load-balanced swarm of VMs with a single front end that can serve in a classroom setting. This can be set up with minimal code and time, either by a savvy instructor or support centers as needed.

2.2 Education Support

The multi-stage and iterative process of biological data analysis is difficult and time consuming. It is also typically necessary to write custom code to visualize biological data or consolidate the results from different analyses. Students with limited coding experience face a steep learning curve that makes the analysis process take excessive amounts of time. As a result, the quality of the analyses can suffer. While coding skills are extremely useful for students to learn, they are often not the focus of their education or academic interests. Recognizing these challenges, we have created a set of interactive tools and tutorials to guide students and researchers through the process of count data analysis (Figure 1).

The successive approach of the tutorials allows us to meet the learners where they are, providing support to extend their understanding without overwhelming them. No coding experience or programming knowledge is necessary to use the freeCount tools or to conduct the analysis workflows described in the tutorials. After running the tutorials, users can immediately explore, manipulate, and visualize their own complex omics data. Students who have an inherent investment in their own data are motivated to achieve a deeper understanding of the analysis. Continued engagement is promoted by the rapid, interactive nature of the process. Students are also better positioned to understand, interpret, and critique publications after interacting with the publication standard visualizations produced by freeCount. The most impactful outcome of our research is the development of a framework that circumvents the typical barriers that students encounter when performing computational analysis, while promoting a deeper understanding of the data and analysis process.

The freeCount tools and tutorials have been used at ND in biology classes with both first year undergraduate students and early graduate students. The undergraduate students were able to learn how to conduct reproducible and high-quality analysis of research data they generated using the Shiny server swarm option. In this

case, the Genomics and Bioinformatics Core Facility managed the VMs on Jetstream2 and provided guest classroom instruction. The graduate student course included numerous teaching assistants for the ten supported sections of the undergraduate course. These graduate students used the local implementation of freeCount to gain expertise in the same data and workflow, creating a pool of expertise supporting the undergraduate course and their own graduate data analysis. Another graduate student used the workflow to perform an individual analysis and complete a dissertation chapter.

3 SIMILAR FRAMEWORKS

Other frameworks exist that support the analysis of count data, such as Galaxy [5], RNAlysis [17], DEBrowser [10], ideal [13], PIVOT [18], NetSeekR [16], and Curare and GenExVis [2]. Each of these frameworks provide tools to perform differential expression analysis, but they vary widely in their additional features and analysis options. Furthermore, many of these frameworks were designed specifically for use with RNA sequencing data. It is also more or less difficult for users to install or access the framework tools. Students who are new to exploratory data analysis can also find some of the interfaces to be complicated and difficult to navigate.

4 ANALYSIS FRAMEWORK

4.1 Overview

The freeCount framework leverages the reactive data processing and visualization components of R Shiny applications to aid researchers in the process of count data analysis. We created separate applications to provide a modular set of tools for the most common analyses. Tutorial and demo data are shipped with each application, which support and guide users through the iterative steps of data assessment, processing, and analysis. Users are also able to export R markdown files with the analysis conducted in each application.

4.2 DE Analysis

As the methods were originally developed for bulk RNAseq data, our modules are named after the commonly performed DE analysis. freeCount DA allows users to identify changes in features using either exact tests or general linearized models (GLMs) with edgeR [4, 14]. Users begin the assessment of the raw count data by visualizing distributions using bar plots, followed by trimmed mean of M-values (TMM) normalization to control for library size bias. Broad sample-level patterns are visualized with ordination plots, including both principal component analysis (PCA) and multi-dimensional scaling (MDS). PCA is commonly used in gene expression analyses and MDS is common for metagenomics. Users can select different pairwise comparisons or provide custom models to test for differces between groups. It is also possible to filter the results for statistical and biological significance by adjusting false discovery rate (FDR) and log2-fold change (LFC) cut offs. The results and thresholds are visualized using mean-difference (MD) plots, volcano plots, and heatmaps. Tables with the resulting DE features are formatted for downstream analysis, such as functional analysis or set operations.

4.3 Network Analysis

freeCount NA guides researchers through the construction of unsigned networks using normalized count data with weighted gene co-expression network analysis (WGCNA) [8, 11]. Appropriately normalized count data can be generated using freeCount DA. Researchers begin assessing their normalized count data by examining the hierarchical clustering of samples. Next, researchers can supervise network construction through the iterative adjustment and assessment of network features. Several reactive images are produced by the application to aid in the assessment, including network topology plots and clustering dendrograms. Results tables with the eigengene data for each module and sets of features associated with each network module can be downloaded and used in downstream analyses.

4.4 Functional Analysis

freeCount FA provides researchers with two different test statistics using topGO [1], Fisher's exact and Kolmogorov-Smirnov (KS) like tests. The Fisher's exact test statistic assumes independence of feature observations and is appropriate for the downstream analysis of DE results. The KS like test assumes dependence and is appropriate for the downstream analysis of network results. The results of the functional analysis are visualized using bar, density, and dot plots. Subgraphs of the significant results may be constructed and downloaded using the application. Results tables with lists of potential biological, molecular, and cellular functions for the sets of input features are also generated. The results tables can be filtered by the unadjusted p-values.

4.5 Set Operations

The freeCount set operations application enables the integration of results from the different analyses. Set operations include subset extractions, in addition to visualizations with Venn and Euler diagrams using ggVennDiagram [6] and eulerr [12]. The set operations can be used by researchers to identify commonly shared or distinct sets of features associated with different analyses. All tables and images generated by the Shiny applications can be downloaded. The images are high-resolution, color blind safe, and publication ready.

4.6 Base Code

There is a distinct trade-off between ease-of-use and flexibility. R Shiny applications constrain the analyses to the reactive elements made available by the application authors. As a result, some options may not be included and there will be parameters that cannot be tuned. To support users who require specific functionality, we provide the base R code required to replicate the tutorials. Users can use the applications and tutorials to learn the analysis process, develop their understanding quickly, replicate the analyses in base R, and then customize the code more easily. This approach helps lower the barriers to entry, then builds the complexity through successive passes with verifiable outputs. Thus, allowing for a smoother learning curve while providing maximum flexibility.

5 SUMMARY & OUTLOOK

Biological data sets are typically high-dimensional and large, often requiring the use of multiple bioinformatic software tools to analyze. Many of the available software tools require the analyst to have some knowledge of coding in a programming language to operate. Furthermore, the analysis of biological data heavily relies on the iterative application of visualization and processing methods. The freeCount tools and tutorials guide students and researchers through the complicated process of biological data analysis without the need for any prior experience with coding.

With freeCount there is no need to freak out. The freeCount tools provide an easy to run, scalable platform for individual researchers or classes of various experience levels to interactively engage with the analysis of count data. Count data was chosen as it encompasses three common workflows (bulk RNAseq, scRNAseq, and metagenomics) that rely heavily on visualization. Count data also requires many repeated steps to properly analyze and processes, which is conducive to automation. Flexible options for deployment are provided to support count data analysis at an individual level. It is also possible to employ a central support center with minimal technical requirements.

freeCount is easy to run, easy to learn, easy to scale. We focused on providing easy to run R Shiny applications that facilitate exploratory data analysis and maintain user engagement. Tutorials are paired with each application to build deeper understanding during the analysis process. To aid research collaboration and reproducibility, users can create R markdown files of the analysis completed with each application. Additionally, R scripts with the code for the analysis methods conducted by each application are provided.

The freeCount analysis framework tools and tutorials are freely available on GitHub (https://github.com/ElizabethBrooks/freeCount). The students and researchers at ND have provided valuable feedback for the continued refinement of the freeCount analysis framework. Further development of the freeCount tutorials and applications are ongoing.

ACKNOWLEDGMENTS

The authors would like to thank the students and researchers at ND who provided feedback for the application tools and tutorials. A special thank you to the Schorey (William McManus) and Pfrender (Neil McAdams, Bret Coggins, Nitin Vincent) labs for feature feedback. This project was supported by Notre Dame Research and the National Science Foundation (NSF) grant "Collaborative Research: EDGE FGT: Genome-wide Knock-out mutant libraries for the microcrustacean Daphnia" (2220695/2324639 to Sen Xu and 2220696 to Michael E. Pfrender). This work used Jetstream2 at Indiana University through allocation BIO230029 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296.

REFERENCES

[1] Adrian Alexa and Jorg Rahnenfuhrer. 2023. topGO: Enrichment Analysis for Gene Ontology. https://doi.org/10.18129/B9.bioc.topGO R package version 2.52.0.

[2] Patrick Blumenkamp, Max Pfister, Sonja Diedrich, Karina Brinkrolf, Sebastian Jaenicke, and Alexander Goesmann. 2024. Curare and GenExVis: a versatile toolkit for analyzing and visualizing RNA-Seq data. *BMC Bioinformatics* 25, 1 (March 2024).

[3] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2024. shiny: Web Application Framework for R. https://shiny.posit.co/ R package version 1.8.1.9000, https://github.com/rstudio/shiny.

[4] Yunshun Chen, Lizhong Chen, Aaron T. L. Lun, Pedro L. Baldoni, and Gordon K. Smyth. 2024. edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. https://doi.org/10.1101/2024.01.21.576131 arXiv:https://www.biorxiv.org/content/early/2024/01/24/2024.01.21.576131.full.pdf

[5] The Galaxy Community. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* 50, W1 (2022), W345–W351. https://doi.org/10.1093/nar/gkac247

[6] Chun-Hui Gao, Chengjie Chen, Turgut Akyol, Adrian Dusa, Guangchuang Yu, Bin Cao, and Peng Cai. 2024. ggVennDiagram: Intuitive Venn diagram software extended. *iMeta* 3 (02 2024). https://doi.org/10.1002/imt2.177

[7] David Y Hancock, Jeremy Fischer, John Michael Lowe, Winona Snapp-Childs, Marlon Pierce, Suresh Marru, J Eric Coulter, Matthew Vaughn, Brian Beck, Nirav Merchant, Edwin Skidmore, and Gwen Jacobs. 2021. Jetstream2: Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced Research Computing* (Boston MA USA). ACM, New York, NY, USA.

[8] Steve Horvath. 2011. *Weighted network analysis* (2011 ed.). Springer, New York, NY.

[9] Wolfgang Huber, Vince Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper Hansen, Rafael Irizarry, Michael Lawrence, Michael Love, James MacDonald, Valerie Obenchain, Andrzej Oleś, and Martin Morgan. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* 12 (01 2015), 115–21. https://doi.org/10.1038/nmeth.3252

[10] Alper Kucukural, Onur Yukselen, Deniz M Ozata, Melissa J Moore, and Manuel Garber. 2019. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* 20, 1 (Jan. 2019), 6.

[11] Peter Langfelder and Steve Horvath. 2012. Fast R Functions For Robust Correlations And Hierarchical Clustering. *Journal of statistical software* 46 (03 2012). https://doi.org/10.18637/jss.v046.i11

[12] Johan Larsson. 2024. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. https://CRAN.R-project.org/package=eulerr R package version 7.0.2.

[13] Federico Marini, Jan Linke, and Harald Binder. 2020. ideal: an R/Bioconductor package for interactive differential expression analysis. *BMC Bioinformatics* 21, 1 (Dec. 2020), 565.

[14] Davis J McCarthy and Gordon K Smyth. 2009. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25, 6 (March 2009), 765–771.

[15] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[16] Himangi Srivastava, Drew Ferrell, and George V Popescu. 2022. NetSeekR: a network analysis pipeline for RNA-Seq time series data. *BMC Bioinformatics* 23, 1 (Jan. 2022), 54.

[17] Guy Teichman, Dror Cohen, Or Ganon, Netta Dunsky, Shachar Shani, Hila Gingold, and Oded Rechavi. 2023. RNAlysis: analyze your RNA sequencing data without writing a single line of code. *BMC Biol.* 21, 1 (April 2023), 74.

[18] Qin Zhu, Stephen A Fisher, Hannah Dueck, Sarah Middleton, Mugdha Khaladkar, and Junhyong Kim. 2018. PIVOT: platform for interactive analysis and visualization of transcriptomics data. *BMC Bioinformatics* 19, 1 (Dec. 2018).

Received 26 April 2024