OXFORD

## Gene expression

# The START App: a web-based RNAseq analysis and visualization resource

Jonathan W. Nelson[1], Jiri Sklenar[1], Anthony P. Barnes[2] and Jessica Minnier[1,3,]*

[1]The Knight Cardiovascular Institute, [2]Department of Pediatrics and [3]School of Public Health, Oregon Health & Science University, Portland, OR 97239-3098, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

## Abstract

**Summary:** Transcriptional profiling using RNA sequencing (RNAseq) has emerged as a powerful methodology to quantify global gene expression patterns in various contexts from single cells to whole tissues. The tremendous amount of data generated by this profiling technology presents a daunting challenge in terms of effectively visualizing and interpreting results. Convenient and intuitive data interfaces are critical for researchers to easily upload, analyze and visualize their RNAseq data. We designed the START (Shiny Transcriptome Analysis Resource Tool) App with these requirements in mind. This application has the power and flexibility to be resident on a local computer or serve as a web-based environment, enabling easy sharing of data between researchers and collaborators.

**Availability and Implementation:** Source Code for the START App is written entirely in R and can be freely available to download at https://github.com/jminnier/STARTapp with the code licensed under GPLv3. It can be launched on any system that has R installed. The START App is also hosted on https://kcvi.shinyapps.io/START for researchers to temporarily upload their data.

**Contact:** minnier@ohsu.edu

## 1 Introduction

Dynamic and stable RNA expression patterns represent the genomic output of complex transcriptional regulatory networks required for proper cellular function. In fact, most cell types can be defined by their transcriptional repertoire and many diseases are characterized by the presence, absence or altered expression of various transcripts (Luna-Zurita and Bruneau, 2013; Sparano *et al.*, 2015). Technologies to accurately quantify RNA expression levels represent an ever-evolving field from Northern blotting to reverse transcription PCR (RT-PCR) and microarrays. The emergence of newer technologies has been driven by the increased throughput of transcriptional data generation. One of the newest and most sensitive technologies for measuring transcript levels is RNA sequencing (RNAseq), an approach that has generated unprecedented amounts of RNA expression data (Han *et al.*, 2015).

The wealth of information contained with these datasets has resulted in a rapid expansion of RNAseq as a technique used by many investigators. This has been accompanied by an explosion in the number of software tools investigators can use to analyze and explore their data such as, easyRNASeq (Delhomme *et al.*, 2012), RNASeqGUI (Russo and Angelini, 2014) and ClustVis (Metsalu and Vilo, 2015) among others (Poplawski *et al.*, 2016). We sought to create a tool in which a wet-bench scientist with little computational programming background would feel comfortable to explore, visualize and share their data. The START App has many advantages over previous tools including offering multiple types of visualizations, publication quality graphics and an intuitive web-based graphical user interface. Furthermore, the START App may be utilized as part of a purely web-based workflow from raw RNA-seq data to results in conjunction with BrowserGenomes.org (Schmid-Burgk and Hornung, 2015) which may eliminate the need for a computational biologist for experiments with straight-forward study design.

It is critical that researchers be able to effectively extract the most information possible from their data given the great deal of

time and resource investment required for such experiments. Software applications that facilitate such efforts may offer additional opportunities for investigation of 'big data' experiments resulting in findings that go beyond the original hypotheses they were collected to address. Therefore, sharing the data with multiple investigators can promote collaboration and add tremendous value to the greater research community.

To address this need, we have designed the START App to provide researchers with increased flexibility to easily upload and visualize RNAseq data. The App visualizes data in multiple ways that will be useful for scientist to understand their data. Critical to facilitating data sharing capabilities, the App can be utilized within a web browser environment for easy access as well as enabling seamless sharing of data between collaborators.

## 2 Methods

The START App is a web-based application written entirely in the open-source R programming language (R Core Team, 2015) using the Shiny framework (Chang *et al.*, 2016). It is fully cross-platform and can be launched locally from any computer that has R installed. Alternatively, users can host their own version of the app with their transcriptome data on a local or remote server so that other users may access the app and data from a website without installing R or the application. The START App has been most extensively tested in a Chrome browser environment. Users may protect their data through firewalls or web-based authentication services. We are also hosting the START app on a public website hosted by the shinyapps.io server (https://kcvi.shinyapps.io/START) where users can explore the features of the app and temporarily upload their data. While we do not anticipate removing the START App from the internet, in compliance with the Bioinformatics policy, we will ensure that the START App is online for 2 years following the publication date. Each data upload to the START website creates a unique instance on the shinyapps.io server, and is therefore only viewable to the user who uploaded the data during a single browser session. There are no limitations to the number of groups that can be analyzed, however, the available RAM is limited to 8 GB which may limit the number of genes (over 5000) that can be simultaneously visualized on the Heatmaps panel. Finally, we strongly encourage users of the START App to familiarize themselves with the terms and services of shinyapps to ensure compliance with any and all restrictions surrounding protected health information (PHI).

## 3 Results

The landing page of the START App contains a *Getting Started* panel with detailed information about how to format and upload a transcriptome dataset. On the *Input Data* panel users may upload transcriptome data consisting of expression values, statistical measures of comparison (such as fold change between groups or a test statistic), and corresponding measures of significance (such as *P*-values or *q*-values), or they may upload 'raw' RNAseq count data generated from read alignment (assuming a straightforward independent group design) that will then be analyzed by the app with a simple normalization and regression analysis pipeline (Law *et al.*, 2014). Additionally, the application is pre-loaded with example data, allowing users to demo analyses and visualizations that the app provides, described below.

The START App provides visualizations for multiple levels of analysis of RNAseq data. These range from comparing transcriptional similarity of experimental groups as a whole in the *Group Plots* panel to querying features of the dataset at the individual transcript level with the *Gene Expression Boxplots* panel. This flexibility allows users to perform either broad group analyses or ask targeted questions about their experimental results. These functionalities are presented in a set of panels in the app:

The *Group Plots* panel contains graphs comparing the similarity or dissimilarity of experimental groups to one another using principal component analysis. Additionally, pairwise comparisons of each sample can be visualized on a sample distance heatmap.

The *Analysis Plots* panel contains scatter plots that allow 2-group comparisons to be either visualized as a volcano plot (log Fold-Change versus log *P*-value) or as an expression scatter plot (logCPM Group 1 versus logCPM Group 2). Both of these scatter plots are content-aware interactive visualizations allowing users to hover their mouse over individual data points to identify the transcript representing the data point.

The *Gene Expression Boxplots* panel contains a box and whiskers visualization and dotplot (stripchart) of transcript expression data that is searchable by transcript identifier. Expression of single or multiple transcripts can be simultaneously visualized as raw counts, log CPM, or the data format inputted by the user.

The *Heatmaps* panel contains a utility for creating a heatmap of transcriptional data. Users can decide which experimental groups they want to compare to drive the organization of the heatmap as well as the number of transcripts that will be visualized. Additionally, users can provide a subset list of transcripts they wish to use to visualize as a heatmap.

Finally, the *Info* panel contains information about how to run the shinyapp locally on a computer which has R installed as well as the shinyapps.io Terms of Service.

It is important to note, however, that proper interpretation of the analysis results and visualizations may require additional statistical and bioinformatics expertise for a given set relational results. We welcome the contribution of users and experts in the future development of the app's source code on github and encourage researchers to copy and modify the code to suit their specific analysis needs. The development of this app is ongoing and we intend to add to and improve upon the visualization and analysis features.

## Acknowledgements

## Funding

## References

Chang,W. *et al.* (2016) shiny: Web Application Framework for R. R package version 0.13.1. http://CRAN.R-project.org/package=shiny

Delhomme,N. *et al.* (2012) easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics*, 28, 2532–2533.

Han,Y. *et al.* (2015) Advanced applications of RNA sequencing and challenges. *Bioinform. Biol. Insights*, 9, 29–46.

Law,C.W. *et al*. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*., **15**, R29.

Luna-Zurita,L. and Bruneau,B.G. (2013) Chromatin modulators as facilitating factors in cellular reprogramming. *Curr. Opin. Genet. Dev*., **23**, 556–561.

Metsalu,T. and Vilo,J. (2015) ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res*., **43**, W566–W570.

Poplawski,A. *et al*. (2016) Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective. *Brief. Bioinform*., **17**, 213–223.

R Core Team. (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Russo,F. and Angelini,C. (2014) RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics*, **30**, 2514–2516.

Schmid-Burgk,J.L. and Hornung,V. (2015) BrowserGenome.org: web-based RNA-seq data analysis and visualization. *Nat. Methods*, **12**, 1001–1001.

Sparano,J.A. *et al*. (2015) Prospective validation of a 21-gene expression assay in breast cancer. *N. Engl. J. Med*., **373**, 2005–2014.