Introduction

RNA-Seq analysis involves carrying out a sequence of steps to analyze, visualize and interpret datasets. The key steps are pre-processing of data, analyzing using statistical techniques such as hypothesis testing, ANOVA, etc., enrichment analysis of differentially expressed gene sets using various biological databases and, identifying hidden patterns using techniques such as weighted gene co-expression network analysis (WGCNA). Each of these steps are facilitated with suitable visualization aids such as principal components visualization, Heatmap, Boxplot of expression levels of genes in different groups, Volcano plot, Venn diagrams and fold change-fold change plot.

Several statistical tools and pipelines exist to carry out RNA-Seq analysis. Most of these tools are licensed. Furthermore, they require users to have a background in programming and statistics making it difficult for biologists with limited statistical and programming knowledge. Motivated to address this problem, there is a sudden spur of research activities in developing intuitive web based graphical interface applications such as *shinyngs* [1], *START* [2], *Degust* [3], *Explore DEG* [4], *DEBrowser* [5]. These applications are user friendly and enabled biologists to perform RNA-Seq data analysis with ease. However, these applications still lack several crucial and useful features in one or more of the steps mentioned earlier. To address these inadequacies, we developed the tool Shiny-Seq that has many new features. As discussed below, these features help to do more comprehensive data analysis and obtain accurate results through suitable statistical analysis and correction techniques.

Shiny-Seq is a comprehensive application tool that comprises of carefully chosen statistical techniques and R-packages to carry out the various steps required to do down-stream differential expression analysis, and, integrating them to function as a coherent workflow with interactive, visualization, analysis and report generation capabilities. The application is embedded with as much data analysis and visualization features as possible. It is designed to act as an interface between the user and the pipeline by taking inputs from the user at each step of the pipeline and calling the appropriate modules in the pipeline for the purpose of exploratory and differential expression analysis of RNA-Seq data. We highlight several important new features of Shiny-Seq.

Biological experimental measurements often contain expression level variabilities due to laboratory conditions, technical factors, and genetic factors, differences in chemicals used persons who conduct experiments or any other un modeled factors and exclusion of some relevant measurement variables [6, 7]. Such variabilities are described as expression heterogeneity [7]. Since failing to take such variabilities into account in statistical analysis can lead do incorrect conclusions, we provide batch effect analysis feature to identify any batch effects in measurements and use discovered effects as inputs in the downstream analysis.

Another important feature that we provide in our tool is to conduct quality control check and support P-value correction while doing hypothesis testing. This step is crucial as we analyze genes using multiple testing techniques. Without such feature, there is a danger of drawing wrong conclusion in the differentially expressed genes.

Enrichment analysis of differentially expressed gene sets using biological databases such as KEGG, Gene Ontology is useful to validate and find relationship with other experimental studies. Besides supporting enrichment analysis, Shiny-Seq also offers prediction of transcription factors and conducting weighted gene co-expression network analysis [8]. WGCNA is useful to identify candidate biomarkers or therapeutic targets. It is important to note that none of the existing applications support enrichment analysis except *DEBrowser* and, prediction of transcription factors and WGCNA are integrated only in Shiny-Seq. To summarize, our application integrates many independent tools and helps the users to conduct a comprehensive analysis using a single tool, thereby, reducing their burden significantly.

Visualization aids help to understand data and analysis results better. Furthermore, they help to make decisions and selecting certain choices during a comprehensive RNA-Seq analysis. In this regard, visualization features that are unique to Shiny-Seq are Venn diagram and fold change-fold change plots.

Summarization of analysis results is an important step and is ignored by all the existing applications. Shiny-Seq offers help to users by collating all key insights and outputs obtained during various stages of the analysis and produce a PowerPoint presentation. To the best of our knowledge, none of the existing applications supports this feature.

Methods

In this section, we provide details regarding components implemented and packages used in the various steps of Shiny-Seq.

Data pre-processing

The first step in the underlying pipeline involves preparation of a count table. The count table, if readily available, is uploaded by the user else the user provides the location of the directory containing the files generated by *Kallisto* [9] (as part of the transcript quantification step). Another input required by the application is the annotation file, a matrix where the rows correspond to the sample names and the columns correspond to the different factors associated with the samples.

Normalization

The package *DESeq2* [10] normalizes the dataset by computing a size factor for each sample. The size factor is calculated by taking the median of ratios of each sample to the reference sample.

Batch effect analysis

Batch effect can be induced either by known variables such as processing groups and dates or unknown variables [7]. In Shiny-Seq, we use the function *removeBatcheffect* in *LIMMA* [11] to account for batch effect due to known variables. On the other hand, we use *SVA* [12] to construct surrogate variables to account for the unknown variables that cause heterogeneity in the data.

Variables known to cause batch effect can be identified by visualizing the principal components of normalized data. If samples belonging to a known or surrogate variable cluster together then the presence of batch effect is detected. We can remove batch effects by passing on the normalized data to *LIMMA* or *SVA* and verify by visualizing the removal effect using PCA plots again.

Although batch removal option is offered, the resulting data is used only for visualization purpose and not used in the downstream analysis. However, we account for any detected batch effect by passing on the batch effect information through annotation table as additional input to *DESeq2*.

Statistical Tests

Shiny-Seq supports pairwise comparison of groups in finding genes that explain differences expressed between the groups. This comparison is done using Wald test. Furthermore, it is useful to find a list of genes that cause differences among the groups and this is facilitated by conducting the ANOVA test. Shiny-Seq uses the package *DESeq2* to conduct these tests.

Quality control (P-value correction)

As emphasized earlier, P-value correction is an important step. The Wald test returns a quantity called Z-statistic that assesses the significance of the differential expression (log2 fold change). The test assumes that the Z-statistic takes a standard normal distribution with zero mean and unit variance. Since P-values are calculated from Z-statistic, we conduct a quality check to validate the assumption by visualizing the histogram of P-values. The shape of the histogram helps to determine whether the assumption is violated. In an event of any violation, it is necessary to make correction and this is done using the package *fdrtool* [13].

Data Exploration Aids

Shiny-Seq supports two important data exploratory features, namely, identification of any clusters in the principal components space and among correlated or co-expressed genes. Since principal component analysis is well-known we describe below only WGCNA, an important feature of this tool.

WGCNA

We use the R package WGCNA [8] to identify clusters of correlated/co-expressed genes. WGCNA takes the pre-processed data table and a list of genes as inputs. Note that batch effect is considered whenever it is present. The output is a table with rows having cluster name, number of genes and hub gene in each cluster. The table is interactive in the sense when the user clicks on columns and download button, all the clusters are downloaded as networks which in turn can be visualized in Cytoscape. Furthermore, Heatmaps of a cluster can be visualized to understand variations in the expression of the genes across all treatment and condition groups.

Other Useful Insights

Our application also provides aids in deriving several other insights. They include identification or association of differentially expressed genes with databases such as transcription factor prediction, user provided marker genes or transcription factors list. Shiny-Seq makes use of the R-package *pcaGopromoter* [14] to predict transcription factors.

Visualization Aids

Shiny-Seq supports several visualization aids. These include aids such as Heatmap and Volcano plots that are commonly available in other applications as well. A heatmap is used to visualize relationships among samples. R package *pheatmap* [15] is used to generate heatmaps of different kinds such as differentially expressed genes, top-1000 genes showing variability and transcription factors. Volcano plot helps to visualize differentially expressed genes obtained from a hypothesis test as a fold-change versus P-value plot in logarithmic scale.

While Heatmaps and volcano plot are useful to visualize hypothesis test results of a single comparison they do not have the capability to compare results obtained from two hypothesis tests. Shiny-Seq addresses this gap through two useful aids that we describe below.

Venn diagram

A Venn diagram helps to visualize the overlap among selected gene lists of differentially expressed genes provided by the user. We use the R package *Venndiagram* [16] to generate this plot.

Fold change Fold change plot

It is an interactive plot that again takes two user selected gene lists as input and plots log2 fold changes of one gene list against the other. Each point in the plot is a differentially expressed gene. The user can hover over points to identify the gene name and observe the differences in log2 fold change values common to the two gene lists.

Generation of report

One of the special features of Shiny-Seq is that it collates all the important outputs generated during each step of the analysis and summarizes the results into a PowerPoint presentation which can then be downloaded and shared with collaborators. These outputs include PCA plots of batch effects removed data, top-10 up-regulated and down-regulated genes, enrichment analysis results, etc. The R package *ReporteRs* [17] is used to generate the PowerPoint presentation.

Usage example

The application tool is useful to carry out analysis of the RNA-Seq data that is obtained from experimental studies conducted on various cell types derived from mice and humans. The source code of this application is written in the open-source R programming language [23] using the Shiny framework [24, 25]. It is a platform independent application and can therefore be launched locally from any computer that has the language R installed [25]. In addition the source code has been put up on github (https://github.com/szenitha/Shiny-Seq) and the application is hosted on a public website hosted by shinyapps.io.server (https://szenitha.shinyapps.io/shiny-seq3/). Due to memory constrains the features such as WGCNA, enrichment of GO terms and PowerPoint generation cannot be used on the public website. Finally, it is important that users familiarize themselves with the terms and services of shinyapps to ensure compliance with any and all restrictions surrounding protected health information [2].

Future work

There are several useful features we intend to add to make the application even more comprehensive. This section summarizes the additional features proposed to be added to improve the visualization and analysis. 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 *Tophat* [18], *HTSeq-counts* [19], *RSEM* [20] and *Sailfish* [21].

References

[1] “shinyngs." https://github.com/pinin4fjords/shinyngs, accessed: 2017-10-28.
[2] J. W. Nelson, J. Sklenar, A. P. Barnes, and J. Minnier, “The start app: a web-based rnaseq analysis and visualization resource," Bioinformatics, p. btw624,2016.
[3] D. R. Powell, “Degust." http://victorian-bioinformatics-consortium.github.io/degust/, accessed: 2017-10-28.
[4] “sushi." http://fgcz-shiny.uzh.ch/fgcz exploreDEG app/, accessed: 2017-03-17.
[5] A. Kucukural and N. Merowsky, “Debrowser documentation," 2016.
[6] http://www.nature.com/nrg/journal/v11/n10/full/nrg2825.html?foxtrotcallback=true
Tackling the widespread and critical impact of batch effects in high-throughput data
Jeffrey T. Leek
[7] http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030161
Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis
Jeffrey T Leek, John D Storey
[8] https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/
[9] C. Robert and M. Watson, “Errors in rna-seq quantification affect genes of relevance to human disease," Genome biology, vol. 16, no. 1, p. 177, 2015.

[10] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for rna-seq data with deseq2." Genome biology, vol. 15, p. 550,
2014.
[11] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K.Smyth,”limma powers differential expression analyses for rna-sequencing and microarray studies." Nucleic acids research, vol. 43, p. e47, Apr. 2015.
[12] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, “The sva package for removing batch effects and other unwanted variation in high-throughput experiments." Bioinformatics (Oxford, England), vol. 28, pp. 882{883, Mar. 2012.
[13] 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.
[14] M. Hansen, T. A. Gerds, O. H. Nielsen, J. B. Seidelin, J. T. Troelsen, and J. Olsen, “pcagopromoter-an r package for biological and regulatory interpretation of principal components in genome-wide gene expression data," PloS one, vol. 7, no. 2, p. e32394, 2012.
[15] R. Kolde, “Heatmap manual." https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf, accessed: 2017-10-28.

[16] H. Chen and P. C. Boutros, “Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r,"BMC bioinformatics, vol. 12, no. 1, p. 35, 2011.
[17] D. Gohel, “ReporteRs package manual to generate PowerPoint presentation."https://cran.r-project.org/web/packages/ReporteRs/ReporteRs.pdf, accessed: 2017-10-28.
[18] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, “Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome biology, vol. 14, p. R36, Apr. 2013.
[19] S. Anders, P. T. Pyl, and W. Huber, “Htseq{a python framework to work with high-throughput sequencing data." Bioinformatics (Oxford, England), vol. 31, pp. 166{169, Jan. 2015.
[20] B. Li and C. N. Dewey, “Rsem: accurate transcript quantification from rna-seq data with or without a reference genome," BMC bioinformatics, vol. 12, no. 1, p. 323, 2011.
[21] R. Patro, S. M. Mount, and C. Kingsford, “Sailfish enables alignment-free iso-form quantification from rna-seq reads using lightweight algorithms," Nature biotechnology, vol. 32, no. 5, pp. 462{464, 2014.
[22] R. C. Team, \R: a language and environment for statistical computing. vienna, austria; 2015," 2015.
[23] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, \shiny: Web application framework for r. r package version 0.13. 0 (2016)."
[24] J. W. Nelson, J. Sklenar, A. P. Barnes, and J. Minnier, \The start app: a web-based rnaseq analysis and visualization resource." Bioinformatics (Oxford, England), vol. 33, pp. 447{449, Feb. 2017.