IRIS-DGE: An integrated RNA-seq data analysis and interpretation system for differential gene expression

Brandon Monier[1,†], Adam McDermaid[2,3,†], Jing Zhao[4,5], Anne Fennell[6], Qin Ma[2,3,6,7,*]

[1]Department of Biology & Microbiology, South Dakota State University, Brookings, SD, USA, [2]Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD, USA, [3]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA, [4]Population Health Group, Sanford Research, Sioux Falls, SD, USA, [5]Department of Internal Medicine, Sanford School of Medicine, University of South Dakota, Vermillion, SD, USA, [6]Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD, USA, [7]BioSNTR, Brookings, SD, USA.

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors

Abstract

Motivation: Next-Generation Sequencing has made available much more large-scale genomic and transcriptomic data. Studies with RNA-sequencing (RNA-seq) data typically involve generation of gene expression profiles that can be further analyzed, many times involving differential gene expression (DGE). This process enables comparison across samples of two or more factor levels. A recurring issue with DGE analyses is the complicated nature of the comparisons to be made, in which a variety of factor combinations, pairwise comparisons, and main or blocked main effects need to be tested.

Results: Here we present a tool called IRIS-DGE, which is a server-based DGE analysis tool developed using Shiny. It provides a straightforward, user-friendly platform for performing comprehensive DGE analysis, and crucial analyses that help design hypotheses and to determine key genomic features. IRIS-DGE integrates the three most commonly used R-based DGE tools to determine differentially expressed genes (DEGs) and includes numerous methods for performing preliminary analysis on user-provided gene expression information. Additionally, this tool integrates a variety of visualizations, in a highly interactive manner, for improved interpretation of preliminary and DGE analyses.

Availability: IRIS-DGE is freely available at http://bmbl.sdstate.edu/IRIS/.

Contact: qin.ma@sdstate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The development of high-throughput technologies, such as RNA-seq, have created large amounts of transcriptomic data that require vast computational resources. One common application of RNA-seq data is the determination of differentially expressed genes (DEGs) through a computational process called Differential Gene Expression (DGE) analysis. DGE analyses allow researchers to investigate which genes are significantly differentially expressed across two or more conditions and can provide a meaningful way to attribute differences in gene expression levels to observed phenotypical differences. In implantation of DGE analysis, substantial algorithms have been developed and optimized, such as DESeq (Anders and Huber, 2012), DESeq2 (Love, et al., 2014), edgeR (Robinson, et al., 2010), limma (Ritchie, et al., 2015),

Cuffdiff (Trapnell, et al., 2012), Cuffdiff2 (Trapnell, et al., 2013), and sleuth (Pimentel, et al., 2017), among many others.

DGE analysis of RNA-seq data typically involves computational experience to not only design the methods and experiment, but also to implement the process using one of the many computing languages. This creates an obstacle for users with limited computational experience who want to analyze the gene expression results from their RNA-seq studies and has led to the increased need to interactivity in DGE analyses and integrated visualization generation of DGE results (Perkel, 2018). While there have been substantial efforts to provide platforms for DGE analysis and visualization of DGE results (Ge, 2017; Goff, et al., 2013; Harshbarger, et al., 2017; McDermaid, et al., 2018; Nelson, et al., 2017; Nueda, et al., 2017; Powell, 2015; Younesy, et al., 2015), numerous pitfalls and bottlenecks persist. Among these pitfalls are the difficulty of experimental design implementation, lack of comprehensive integrated preliminary analyses and DGE tools, and lack of functionalities and interactivity related to visualizing the analysis results.

To address these bottlenecks, we have created IRIS-DGE, which is an **I**nteractive **R**NA-seq data analysis & **I**nterpretation **S**ystem for **D**ifferential **G**ene **E**xpression analysis. IRIS-DGE provides a user-friendly platform to analyze gene expression profiles comprehensively and to visualize the results from these analyses using interactive figures and tables. Compared with other DGE analysis and visualization platforms, IRIS-DGE provides a more comprehensive user experience on multiple levels. Key areas of emphasis that IRIS-DGE outperforms other tools are the number of integrated DGE tools, variety of visualizations, useful preliminary and DGE analyses, and features aimed at improving the efficiency, applicability, and widespread use (Figure 1A). Focusing on these four areas, IRIS-DGE outperforms comparable tools, with integration of three of the most commonly used R-based DGE tools, seven differential

visualizations, five distinct and useful analyses, interactivity, and both experimental design options and capabilities for interpreting user-provided design matrices. This tool is capable of performing expression quality check and preliminary analyses and integrates three of the most common R-based DGE tools (*Supplementary Materials S1*), *DESeq2*, *edgeR*, and *limma*, to perform DGE analysis. IRIS-DGE also provides users with a choice of intuitive experimental design options, as well as the ability for users to upload their own design matrix. In addition to the analyses, IRIS-DGE provides numerous interactive visualizations of each analysis, enabling users to gain a more global view of their data and results (Figure 1 B-E). Each of these visualizations is also available for download as a static image plot for inclusion in reports or publications.

IRIS-DGE is freely available for use at http://bmbl.sdstate.edu/IRIS/ and can be loaded through R on a local computer. More information regarding implementation and accessibility can be found in *Supplementary Materials S2*.

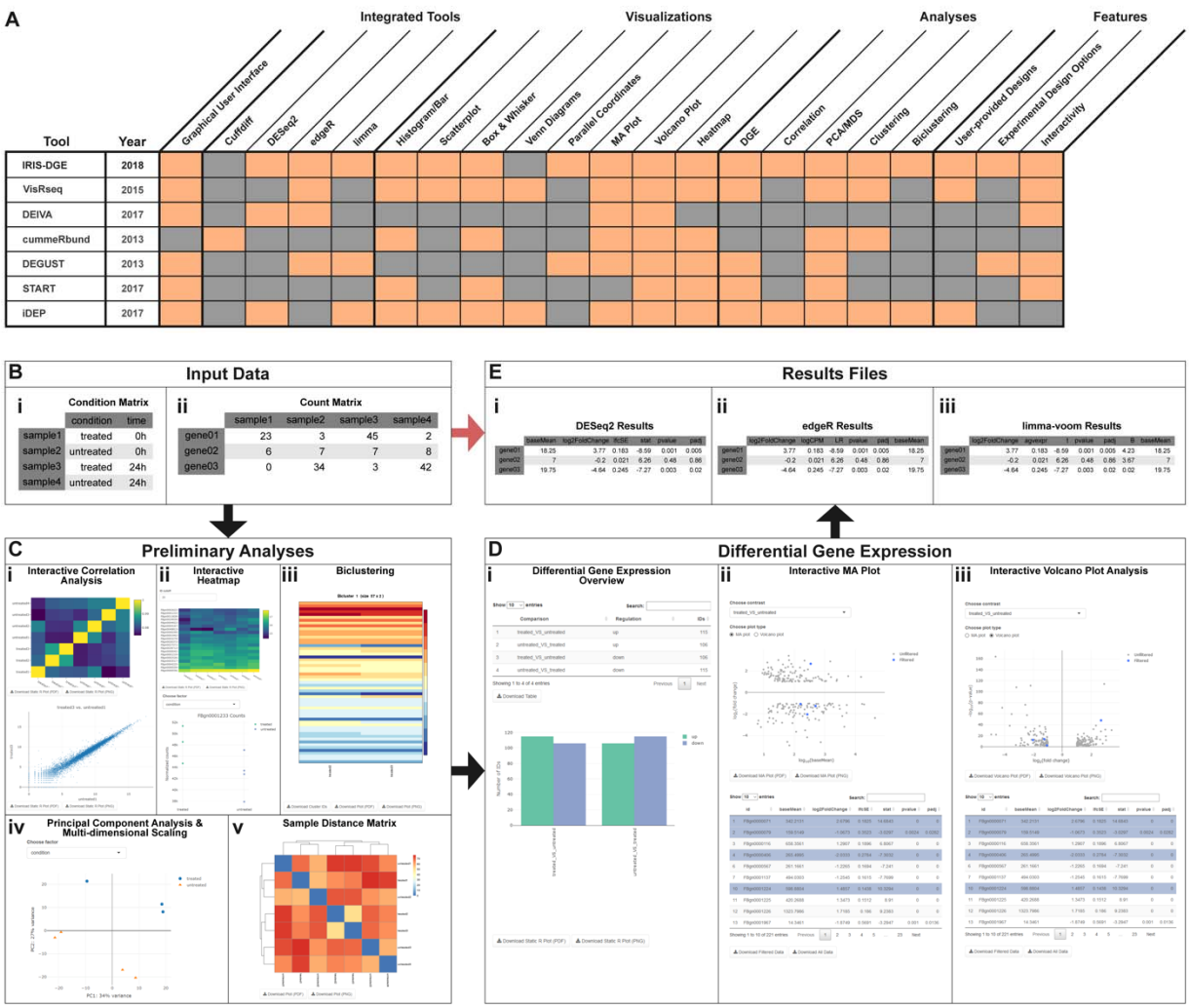


Figure 1: (A) Comparison of IRIS-DGE and six other DGE analysis and visualization tools displaying key novelties of IRIS-DGE, including numerous visualizations and analyses, integrated tools, interactivity, and experimental design capabilities; (B) Required input information for IRIS-DGE: (i) Condition Matrix indicating factor levels for each sample, and (ii) Count Matrix consisting of gene expression levels for each sample, with corresponding sample IDs matching those in the condition matrix; (C) Preliminary Analyses conducted by IRIS-DGE using the Condition and Count matrices, including (i) Interactive Correlation Analysis with pairwise expression scatterplot, (ii)Interactive Heatmap with parallel coordinate plot, (iii) Biclustering, (iv) Principal Component Analysis and Multi-dimensional Scaling, and (v) Sample Distance Matrix with clustering dendrogram; (D) Integrated Differential Gene Expression analysis with visualizations: (i) Differential Gene Expression Overview with table and bar

charts corresponding to up- and down-regulated gene counts, (ii) Interactive MA Plot with DGE results table, and (iii) Interactive Volcano Plot with DGE results table; and (E) Final DGE results extraction with table corresponding to the selected DGE tool.

2 Methods

Required Inputs

IRIS-DGE requires two user-provided input files (Figure 1B and *Supplementary Materials S3*): (1) a gene expression estimation matrix (EEM, also referred to as sample count data) and (2) a condition matrix with factor levels corresponding to the provided samples in the EEM. Once users have uploaded their required data, IRIS-DGE provides two distinct pathways for analysis. Users can choose an in-depth approach utilizing all features of IRIS-DGE, including quality control and preliminary analyses before continuing to DGE analysis. While users are suggested to fully utilize all IRIS-DGE features, we also provide a method for data to be analyzed much more quickly through IRIS-DGE. This approach is referred to as the Expedited Analysis approach and only involves data input and final results extraction.

Expedited Analysis

To focus on the user-friendliness of IRIS-DGE, the tool is designed for users to proceed from EEM to final DGE results extraction in fewer than ten clicks and around one minute through the Expedited Analysis approach (*Supplementary Materials S4).* All inputs, including parameters, specifications, and tools, have a default setting recommended by the authors for best general performance and to assist users with limited computational and analysis experience. Users

would only need to submit their data and indicate their preferred comparisons. This method is provided in order to allow users to quickly generate DGE results without having to go through each of the features included in IRIS-DGE and to extract the usable results from the default parameterizations. The expedited analysis pathway takes users from data input (Figure 1B) directly to DGE results extraction (Figure 1E).

In-depth Analysis

Users who prefer a more investigative approach of their gene expression data could instead utilize all features of IRIS-DGE thorugh the In-depth Analysis pathway, fully exploring the provided quality control and preliminary analyses before moving on to the specific details of their DGE analysis (*Supplementary Materials S5.1*).

Quality Control

After upload and submission, the two input files are first analyzed by IRIS-DGE quality control (*Supplementary Materials S5.2)*. This process involves a sample of the input data provided by the user and boxplots and histograms of read count distributions and a histogram of total read counts by the sample. The purpose of the quality control process is to enable exploration of the submitted data and to verify that there are no unexpected or unexplainable abnormalities in the data, such as low total read counts or individual samples displaying strange distribution behavior.

Preliminary Analyses

Another of the important features of IRIS-DGE is the preliminary analyses (Figure 1C), which can assist users in visualizing their EEM input information and to discover trends in their data

that may provide addition hypotheses to investigate with downstream analyses, including DGE analysis. Preliminary analyses performed by IRIS-DGE include sample correlation analysis and pairwise expression scatterplots (Figure 1Ci), expression heatmaps (Figure 1Cii), biclustering (Figure 1Ciii), principal component analysis and multidimensional scaling (Figure 1Civ), and sample distance matrix with clustering (Figure 1Cv). The figures generated through the preliminary analysis feature of IRIS-DGE are provided in an interactive manner, which allows users to select specific samples or pairwise comparisons to further evaluate. One such example is with the sample correlation analysis and pairwise scatterplots shown in Figure 1Ci. Users can select one cell of the sample correlation matrix corresponding to a comparison between two samples. This will display the pairwise scatterplot for that specific comparison. The user can then scroll over the scatterplot, which will display the gene ID for the indicated data point. More details regarding each preliminary analysis and their respective interactivity can be found in the *Supplementary Materials S5.3*.

Differential Gene Expression Analysis

After submitting and exploring the data, users can move onto the DGE phase of IRIS-DGE (*Supplementary Materials S5.4*). This analysis is performed using any one of the three provided tools: *DESeq2* (Anders and Huber, 2012), *edgeR* (Robinson, et al., 2010), and *limma* (Ritchie, et al., 2015). The default tool is *DESeq2* based on our RNA-seq experience and independent evidence supporting its performance (Sahraeian, et al., 2017), but users can also select one of these other two tools based on their own preference. While there are other high-performing, commonly-used DGE tools available, their compatibility with IRIS-DGE is what excludes their use in IRIS-DGE. Tools like *Cuffdiff* (Trapnell, et al., 2012) that are not R-based or tools like

sleuth (Pimentel, et al., 2017) which do not operate on sample count data are not included due to compatibility issues.

In addition to the DGE tool, the experimental design can also be specified by the user. The designs provided in IRIS-DGE include two-group comparisons for analysis of selected pairwise comparisons, multiple factorial comparisons, classic interaction design, additive models for pairing or blocking of data, main effect testing, and blocked main effect testing. Additionally, IRIS-DGE provides a method for users to specific their own experimental, for the instances when the user needs a design not already included in IRIS-DGE. Each of these methods has unique parameters to specify by the user, typically including which factors are intended for analysis and which specific comparisons are required. After analyzing the data, IRIS-DGE provides an overview displaying the number of up- and down-regulated IDs for each indicated comparison, along with a histogram displaying this information (Figure 1Di). The results table is also available through IRIS-DGE, along with interactive MA (Figure 1Dii) and Volcano plots (Figure 1Diii).

Similar to the figures generated in the Preliminary Analysis section of IRIS-DGE, the plots in the DGE section are also interactive. This feature allows users to gain more specific information from their plots, including highlighting individual or regions of data points on the plot. Doing so will highlight the corresponding row of the DGE results table, showing users which information about genes that may be outliers or fall within a certain region. Conversely, users can select specific gene IDs from the results table, which highlights the location of that gene ID's or set of gene IDs' data points on the corresponding plot. Users who are focused on a specific gene or set of genes can use this feature to easily determine the relative location of their genes in the plot.

IRIS-DGE Outputs

IRIS-DGE provides users with methods for extracting content based on preliminary and DGE analyses. All figures in the QC, Preliminary Analysis, and DGE Analysis sections have the option for users to download as a static image in PDF or PNG format. Additionally, all tables in the DGE Analysis section are downloadable as CSV files, with the final results table being downloaded in its entirety or filtered based on user-provided or default adjusted p-value and log fold-change cutoffs. As part of the biclustering analysis, users can also download a list of gene IDs contained within the specified cluster.

3 Conclusions

IRIS-DGE is a tool developed for comprehensive DGE analysis designed to address current issues with tools intended to perform DGE analysis and visualize and interpret the results. This tool implements numerous features including EEM quality control, preliminary analysis, and DGE utilizing the most commonly used R-based DGE tools in a user-friendly, comprehensive platform. It is noteworthy that IRIS-DGE provides advanced experimental design options in an intuitive format, while also allowing users to provide their own design matrix to facilitate efficient DGE analysis for a broad spectrum of users. Each analysis section within IRIS-DGE provides relevant information in a highly-interactive visual format. In addition to providing a comprehensive method for performing DGE analysis through the quality control and preliminary analyses, IRIS-DGE also provides a straightforward method for expedited analysis that allows users to extract DGE results in a very limited amount of time using the default DGE parameters.

It is our belief that this tool will assist users of all computational experience levels and with all DGE requirements.

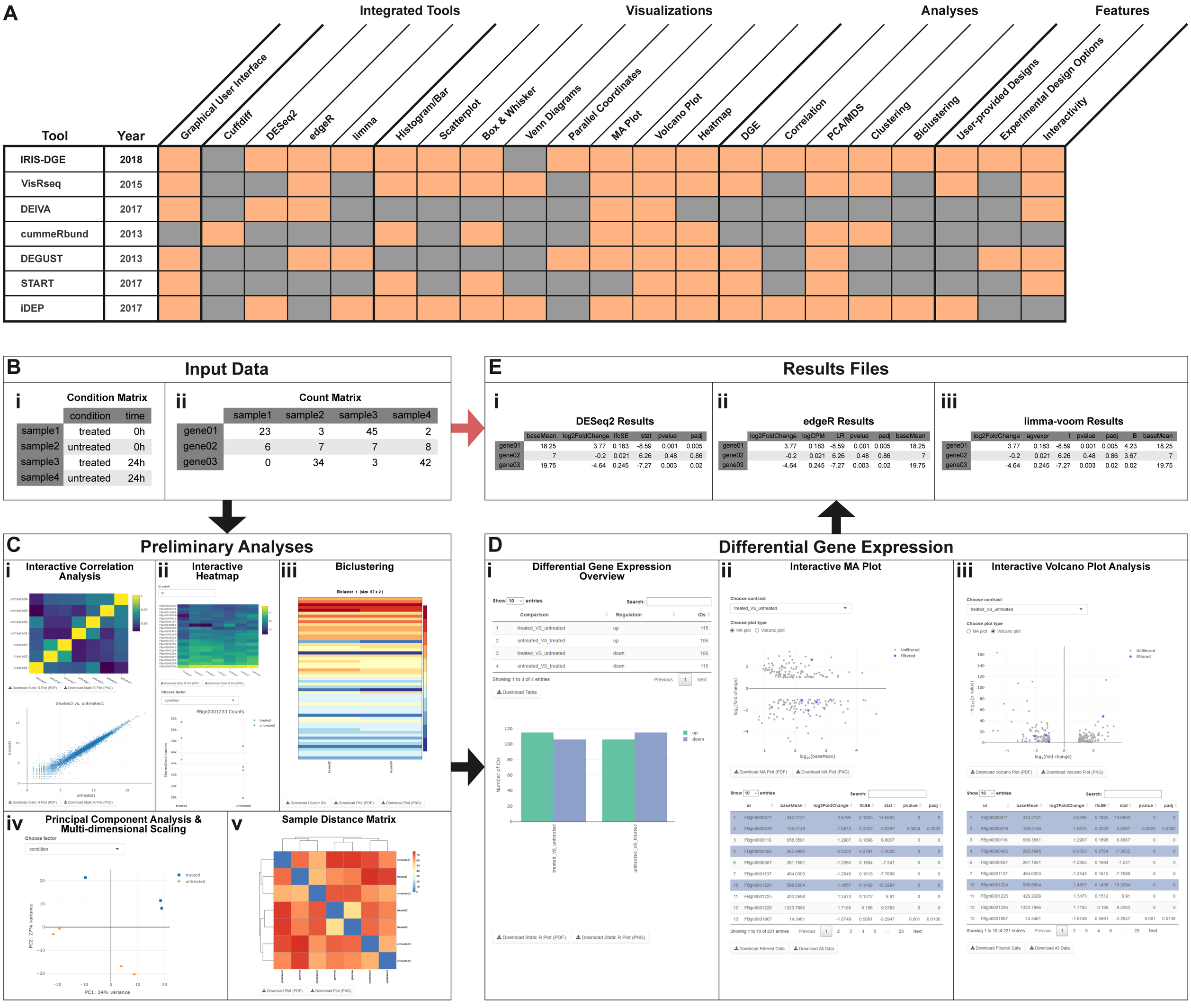
Funding

This work was supported by National Science Foundation / EPSCoR Award No. IIA-1355423, the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University (SDSU). Support for this project was also provided by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number **5P20GM121341** and Sanford Health–SDSU Collaborative Research Seed Grant Program. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (grant number ACI-1548562). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4 References

Anders, S. and Huber, W. Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* 2012.
Ge, S.X. iDEP: An integrated web application for differential expression and pathway analysis. *bioRxiv* 2017.
Goff, L., Trapnell, C. and Kelley, D. cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. *R package version* 2013;2(0).
Harshbarger, J., Kratz, A. and Carninci, P. DEIVA: a web application for interactive visual analysis of differential gene expression profiles. *BMC genomics* 2017;18(1):47.
Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014;15(12):550.
McDermaid, A., *et al.* ViDGER: An R package for integrative interpretation of differential gene expression results of RNA-seq data. *bioRxiv* 2018.
Nelson, J.W., *et al.* The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics* 2017;33(3):447-449.
Nueda, M.J., *et al.* Identification and visualization of differential isoform expression in RNA-seq time series. *Bioinformatics* 2017.

Perkel, J.M. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* 2018;554(7690):133-134.
Pimentel, H., *et al.* Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nature methods* 2017.
Powell, D. Degust: Visualize, explore and appreciate RNA-seq differential gene-expression data. In, *COMBINE RNA-seq workshop*. 2015.
Ritchie, M.E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 2015;43(7):e47.
Robinson, M.D., McCarthy, D.J. and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-140.
Sahraeian, S.M.E., *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature communications* 2017;8(1):59.
Trapnell, C., *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* 2013;31(1):46.
Trapnell, C., *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 2012;7(3):562-578.
Younesy, H., *et al.* VisRseq: R-based visual framework for analysis of sequencing data. *BMC bioinformatics* 2015;16(11):S2.


A
Integrated Tools
Visualizations
Analyses
Features
Tool
Year
Graphical User Interface
Cuffdiff
DESeq2
edgeR
limma
Histogram/Bar
Scatterplot
Box & Whisker
Venn Diagrams
Parallel Coordinates
MA Plot
Volcano Plot
Heatmap
DGE
Correlation
PCA/MDS
Clustering
Biclustering
User-provided Designs
Experimental Design Options
Interactivity
IRIS-DGE 2018
VisRseq 2015
DEIVA 2017
cummeRbund 2013
DEGUST 2013
START 2017
iDEP 2017
B
Input Data
i
Condition Matrix
condition time
sample1 treated 0h
sample2 untreated 0h
sample3 treated 24h
sample4 untreated 24h
ii
Count Matrix
sample1 sample2 sample3 sample4
gene01 23 3 45 2
gene02 6 7 7 8
gene03 0 34 3 42
E
Results Files
i
DESeq2 Results
baseMean log2FoldChange lfcSE stat pvalue padj
gene01 18.25 3.77 0.183 -8.59 0.001 0.005
gene02 7 -0.2 0.021 6.26 0.48 0.86
gene03 19.75 -4.64 0.245 -7.27 0.003 0.02
ii
edgeR Results
log2FoldChange logCPM LR pvalue padj baseMean
gene01 3.77 0.183 -8.59 0.001 0.005 18.25
gene02 -0.2 0.021 6.26 0.48 0.86 7
gene03 -4.64 0.245 -7.27 0.003 0.02 19.75
iii
limma-voom Results
log2FoldChange agvexpr t pvalue padj B baseMean
gene01 3.77 0.183 -8.59 0.001 0.005 4.23 18.25
gene02 -0.2 0.021 6.26 0.48 0.86 3.67 7
gene03 -4.64 0.245 -7.27 0.003 0.02 0.02 19.75
C
Preliminary Analyses
i
Interactive Correlation Analysis
Download Static R Plot (PDF)
Download Static R Plot (PNG)
treated3 vs. untreated1
ii
Interactive Heatmap
Choose factor
condition
FBgn0001233 Counts
treated
untreated
iii
Biclustering
Bicluster 1 (size 57 x 2)
Download Cluster IDs
Download Plot (PDF)
Download Plot (PNG)
iv
Principal Component Analysis & Multi-dimensional Scaling
Choose factor
condition
treated
untreated
PC2: 27% variance
PC1: 34% variance
v
Sample Distance Matrix
Download Plot (PDF)
Download Plot (PNG)
D
Differential Gene Expression
i
Differential Gene Expression Overview
Show 10 entries
Search:
Comparison Regulation IDs
1 treated_VS_untreated up 115
2 untreated_VS_treated up 106
3 treated_VS_untreated down 106
4 untreated_VS_treated down 115
Showing 1 to 4 of 4 entries
Previous 1 Next
Download Table
up
down
Number of IDs
treated_VS_untreated
untreated_VS_treated
Download Static R Plot (PDF)
Download Static R Plot (PNG)
ii
Interactive MA Plot
Choose contrast
treated_VS_untreated
Choose plot type
MA plot
Volcano plot
Unfiltered
Filtered
log2(fold change)
log10(baseMean)
Download MA Plot (PDF)
Download MA Plot (PNG)
Show 10 entries
Search:
id baseMean log2FoldChange lfcSE stat pvalue padj
1 FBgn0000071 342.2131 2.6796 0.1825 14.6843 0 0
2 FBgn0000079 159.5149 -1.0673 0.3523 -3.0297 0.0024 0.0282
3 FBgn0000116 658.3561 1.2907 0.1896 6.8067 0 0
4 FBgn0000406 265.4995 -2.0333 0.2784 -7.3032 0 0
6 FBgn0000567 261.1661 -1.2265 0.1694 -7.241 0 0
7 FBgn0001137 494.0303 -1.2545 0.1615 -7.7699 0 0
10 FBgn0001224 598.8804 1.4857 0.1438 10.3294 0 0
11 FBgn0001225 420.2688 1.3473 0.1512 8.91 0 0
12 FBgn0001226 1323.7986 1.7185 0.186 9.2383 0 0
13 FBgn0001967 14.3461 -1.8749 0.5691 -3.2947 0.001 0.0136
Showing 1 to 10 of 221 entries
Previous 1 2 3 4 5 ... 23 Next
Download Filtered Data
Download All Data
iii
Interactive Volcano Plot Analysis
Choose contrast
treated_VS_untreated
Choose plot type
MA plot
Volcano plot
Unfiltered
Filtered
-log10(p-value)
log2(fold change)
Download Volcano Plot (PDF)
Download Volcano Plot (PNG)
Show 10 entries
Search:
id baseMean log2FoldChange lfcSE stat pvalue padj
1 FBgn0000071 342.2131 2.6796 0.1825 14.6843 0 0
2 FBgn0000079 159.5149 -1.0673 0.3523 -3.0297 0.0024 0.0282
3 FBgn0000116 658.3561 1.2907 0.1896 6.8067 0 0
4 FBgn0000406 265.4995 -2.0333 0.2784 -7.3032 0 0
6 FBgn0000567 261.1661 -1.2265 0.1694 -7.241 0 0
7 FBgn0001137 494.0303 -1.2545 0.1615 -7.7699 0 0
10 FBgn0001224 598.8804 1.4857 0.1438 10.3294 0 0
11 FBgn0001225 420.2688 1.3473 0.1512 8.91 0 0
12 FBgn0001226 1323.7986 1.7185 0.186 9.2383 0 0
13 FBgn0001967 14.3461 -1.8749 0.5691 -3.2947 0.001 0.0136
Showing 1 to 10 of 221 entries
Previous 1 2 3 4 5 ... 23 Next
Download Filtered Data
Download All Data