Interdisciplinary Sciences: Computational Life Sciences (2020) 12:349–354
https://doi.org/10.1007/s12539-020-00383-7

ORIGINAL RESEARCH ARTICLE

Shiny-DEG: A Web Application to Analyze and Visualize Differentially Expressed Genes in RNA-seq

Sufang Wang[1] · Yu Zhang[2,3] · Congzhan Hu[1] · Nu Zhang[1] · Michael Gribskov[4,5] · Hui Yang[1]

Received: 7 April 2020 / Revised: 8 June 2020 / Accepted: 1 July 2020 / Published online: 14 July 2020

Abstract

RNA-seq analysis has become one of the most widely used methods for biological and medical experiments, aiming to identify differentially expressed genes at a large scale. However, due to lack of programming skills and statistical background, it is difficult for biologists including faculty and students to fully understand what the RNA-seq results are and how to interpret them. In recent years, even though, there are several programs or websites that assist researchers to analyze and visualize NGS results, they have several limitations. Therefore, Shiny-DEG, a web application that facilitates the exploration and visualization of differentially expressed genes from RNA-seq, was developed. It integrates multi-factor design experiments, allows users to modify the parameters interactively according to experiments purpose and all analysis results can be downloaded directly, aiming to further assisting the interpretation and explanation of the biological questions. Therefore, it serves better for biologists without programming skills. Overall, this project is of great significance to reveal the mechanism of transcriptome differences.

Keywords RNA-seq · Differentially expressed genes · Shiny · Data visualization

1 Introduction

With the fast development of next-generation sequencing technology (NGS), RNA-seq is one of the most important methods to study gene expression at the transcriptome level [1]. At present, a large number of scientific studies have discovered that human diseases, drug efficacy, plant resistance, etc., are regulated by gene expression [2–4]. Therefore, using RNA-seq to reveal the mechanism is becoming common in biological labs [5–7].

However, due to lack of programming skills and statistical background, it is difficult for biologists including faculty and students to fully understand what the RNA-seq results are and how to interpret them. In recent years, even though, there are several programs or websites that assist researchers to analyze and visualize NGS results, including GSCALite [8], EDGE [9], iDINGO [10], iSeq [11], methylGSA [12], MicroScope [13], RIVET [14], ScanGEO [15], ShinyCNV [16], ShinySISPA [17], STARTApp [18], BLASTmap [19], DiNAR [20], HTPmod [21],Shiny-phyloseq [22], DAME [23], Shiny-Seq [24] and Docker4Circ [25], they have provided good data analysis and visualization experience. However, they have several limitations (Supplemental Table 1). For example, (1) some are designed for experienced R users, which requires programming skills or bioinformatics background; (2) the results cannot be downloaded from websites, which brings difficulties for obtaining the figures or tables; (3) some platforms do not provide parameter selections, which produces the same results no matter what the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12539-020-00383-7) contains supplementary material, which is available to authorized users.

✉ Sufang Wang
sufangwang@nwpu.edu.cn

✉ Yu Zhang
zhangyu@nwpu.edu.cn

1 School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China

2 School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China

3 School of Computer Science and IT, RMIT University, Melbourne, VIC 3000, Australia

4 Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

5 Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA

experiments are, then leading to insufficient data mining; (4) more importantly, none of them could analyze multi-factor design experiments, which does not allow multi-groups comparisons.

Based on this, this project, using the web page programming language, shiny, builds a web page platform to identify DEG in RNA-seq. Through the establishment of a data visualization model and integration of multi-factor design, including the analysis of differential expression genes, clustering analysis and principal component analysis (PCA), users can get better experimental results. At the same time, this platform also provides parameter selection and users can adjust the settings according to different experimental purposes. Therefore, it serves better for biologists without programming skills. Overall, this project is of great significance to reveal the mechanism of transcriptome differences.

2 Materials and Methods

2.1 Programming Language

Shiny-DEG is implemented in R as a Shiny application. The source code could be downloaded from github: https://github.com/sufangwang-npu/shiny-DEG.Users could use Shiny-DEG by two steps: (1) download the source code and (2) launch the package in R or RStudio locally. Shiny-DEG consists of four sections: (1) data upload, (2) data analysis, (3) data visualization and (4) data download. In the following, details in each step will be explained.

2.2 Data Upload

Data can be uploaded as a text file that contains the gene expression level, also called the count table, which is the universal and common file generated by most of the alignment and quantification programs. Shiny-DEG also provides a build-in example data, users could explore the app's features with the example data by clicking on the associated-tabs.

2.3 Data ANALYSIS

When the count table is uploaded, the data is then analyzed by DESeq 2 package to identify differentially expressed genes. Shiny-DEG provided two statistical models: (1) one-factor (default) model which is commonly used to compare control and experimental groups (2) multi-factor model which is one of the most important features in Shiny-DEG.

The one-factor model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where i is group (control or experiment), j is replicated (default is 3).

The multi-factor model is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where i is factor 1, j is factor 2, k is replicated (default is 3). This model is useful when users have two factors and would like to identify DEG in all conditions.

2.4 Data Visualization and Exploration

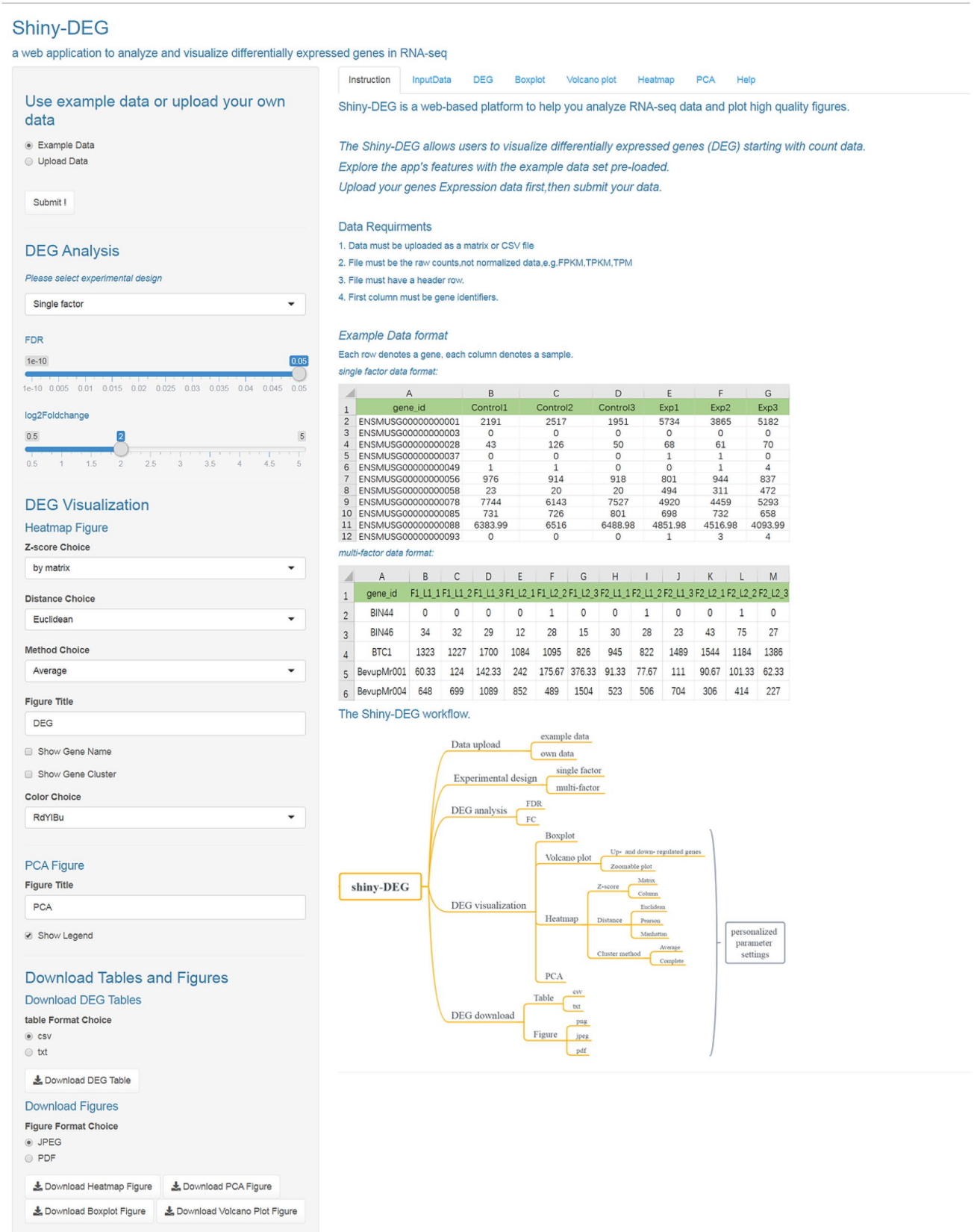
After the count table was analyzed and differentially expressed genes were displayed in the following ways: (1) a DEG table which contains each gene name, basemean, $\log_2$Foldchange ($\log_2$FC) and False Discovery Rate (FDR); (2) Boxplots of gene expression which shows the distribution of gene expression across all samples; ((3) Volcano plots of DEG which shows up-and down-regulated genes in up-left and up-right corners; (4) A Heatmap of cluster analysis which displays the similarity of DEG expression pattern across all samples; (5) Principal Component Analyses (PCA) plots of DEG which reduces the data dimension and captures the most of variances into first two principal components.

2.5 Parameter Choices and Settings

Shiny-DEG allows users to modify several parameters, to better answer and reflect the biological questions. (1) Experimental design, one-factor (default) design or multi-factor design; (2) DEG filter, Shiny-DEG uses both Foldchange and FDR to filter DEG, FDR ranges from 10^{-10} to 0.05 (default is 0.05), Log_2FC ranges from 0.5 to 5 (default is 2); (3) Hierarchical cluster analysis, is performed on the z scores. The key settings including the Z score choices (by matrix or by column), distance choices (Euclidean, Manhattan and Pearson correlation) and cluster methods choices (average or Complete), could be modified. Users could change them depending on experimental purposes; (4) Dendrogram of the cluster, the default is showing the column dendrogram, but users could also choose to display the gene cluster which corresponds to row dendrogram.

2.6 Data Download

All tables and figures generated from Shiny-DEG could be directly downloaded from the website for free. It supports



	A	B	C	D	E	F	G
1	gene_id	Control1	Control2	Control3	Exp1	Exp2	Exp3
2	ENSMUSG00000000001	2191	2517	1951	5734	3865	5182
3	ENSMUSG00000000003	0	0	0	0	0	0
4	ENSMUSG00000000028	43	126	50	68	61	70
5	ENSMUSG00000000037	0	0	0	1	1	0
6	ENSMUSG00000000049	1	1	0	0	1	4
7	ENSMUSG00000000056	976	914	918	801	944	837
8	ENSMUSG00000000058	23	20	20	494	311	472
9	ENSMUSG00000000078	7744	6143	7527	4920	4459	5293
10	ENSMUSG00000000085	731	726	801	698	732	658
11	ENSMUSG00000000088	6383.99	6516	6488.98	4851.98	4516.98	4093.99
12	ENSMUSG00000000093	0	0	0	1	3	4

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	gene_id	F1_L1_1	F1_L1_2	F1_L1_3	F1_L2_1	F1_L2_2	F1_L2_3	F2_L1_1	F2_L1_2	F2_L1_3	F2_L2_1	F2_L2_2	F2_L2_3
2	BIN44	0	0	0	0	1	0	0	1	0	0	1	0
3	BIN46	34	32	29	12	28	15	30	28	23	43	75	27
4	BTC1	1323	1227	1700	1084	1095	826	945	822	1489	1544	1184	1386
5	BevupMr001	60.33	124	142.33	242	175.67	376.33	91.33	77.67	111	90.67	101.33	62.33
6	BevupMr004	648	699	1089	852	489	1504	523	506	704	306	414	227

Fig. 1 The layout of Shiny-DEG

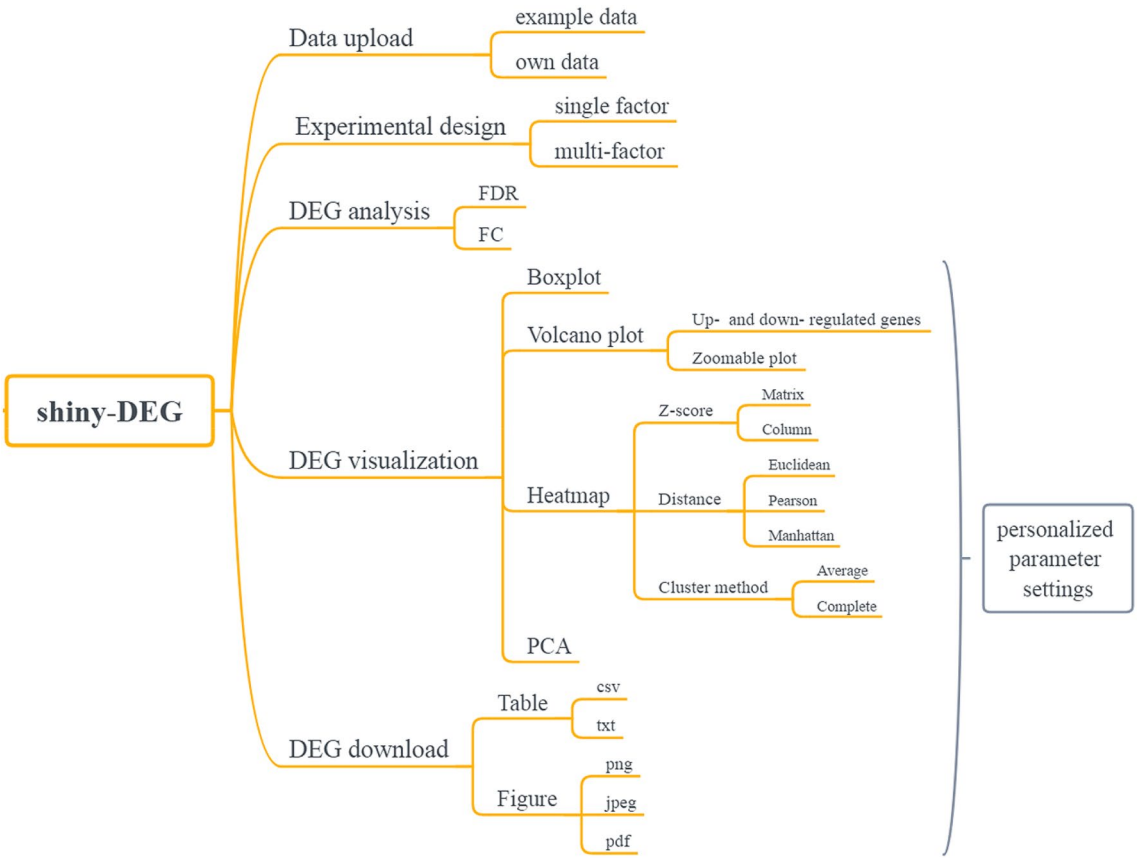


Fig. 2 Workflow of Shiny-DEG

two formats for tables (csv or txt), and two formats for figures with high qualities (jpeg or pdf).

3 Results and Discussion

3.1 The Design and Workflow of Shiny-DEG

Shiny-DEG is designed and programmed by R language. The layout of the webpage is mainly three panels (Fig. 1). Panel 1 is the title and brief explanation of Shiny-DEG. Panel 2 is the sidebar of parameter settings. Panel 3 is the multi-tabs displaying results (tables and figures). In panel 2, there are four parts, (1) Upload data; (2) DEG analysis; (3) DEG visualization and (4) Download part. In panel 3, there are 8 multi-window tabs, including Instruction, InputData, DEG, Boxplot, Volcano plot, Heatmap, PCA and Help.

The workflow of Shiny-DEG is illustrated as follows (Fig. 2): (1) upload expression data; (2) choose an experimental design; (3) choose DEG threshold; (4) DEG visualization; (5) DEG exploration by parameter according to the experimental purpose and (6) DEG download.

The most important two features of Shiny-DEG are: (1) it allows users to explore the data interactively and shows the results according to personalized settings, which may better reflect the biological questions or phenomenon; (2) it has a multi-factor design model, which fits better to experiments with more than one factor and would like to identify DEG across all conditions. Shiny-DEG also provides one example of data with multi-factor design. Users could explore the features and results with the example data.

3.2 The Use and Validation of Shiny-DEG

To validate Shiny-DEG, we chose a gold standard data that mainly focused on differences in gene expression, splicing and RNA editing between embryonic and adult cerebral cortex [26]. We downloaded its gene expression data and uploaded into Shiny-DEG. In this dataset, the authors first considered overall gene expression for transcripts and were able to completely separate the embryonic and adult mice. Through Shiny-DEG, we produced the same figure (Fig. 3), which proved the accuracy of our tool. Then we compared differentially expressed genes by volcano plot (Fig. 3), again,

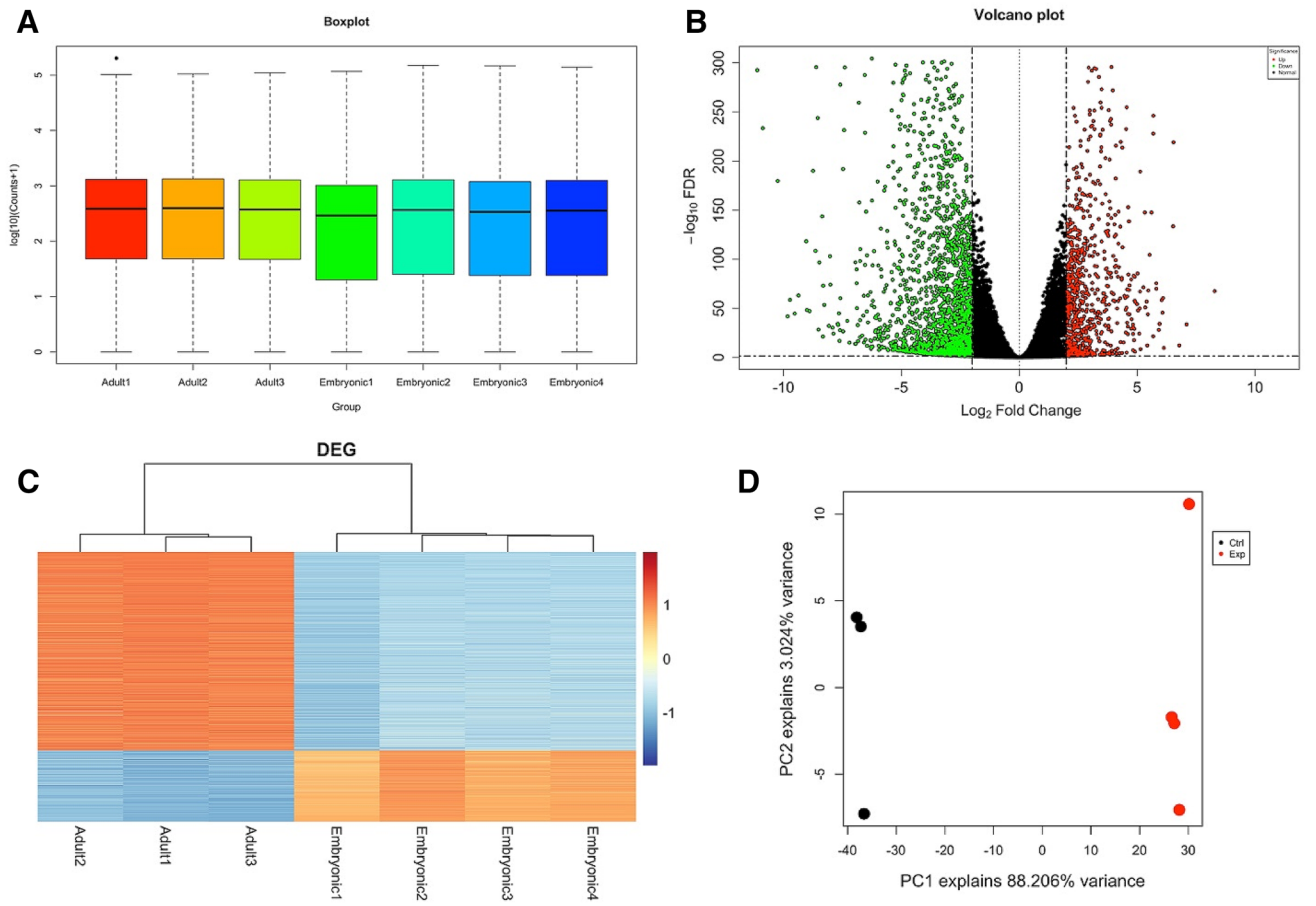


Fig. 3 Validation of the RNA-seq data. **a** Boxplot of gene expression. **b** Volcano plot of the genes. Green dots means down-regulated genes, red dots means up-regulated genes. **c** Heatmap of cluster analysis. **d** Principal component analysis of all samples

Shiny-DEG produced the same plot, which the gene expression pattern is consistent. Therefore, Shiny-DEG is a useful tool for researchers to confirm or generate, hypotheses related to gene expression.

Overall, Shiny-DEG aims to provide a better data analysis and visualization experiences for biological researchers with limited programming skills and bioinformatics knowledge background. However, it has one limitation, which Shiny-DEG did not provide the downstream analysis such as gene ontology enrichment and pathway analysis. In the future, we will keep updating and developing Shiny-DEG, which integrates more analysis and functions.

4 Conclusions

In this research, Shiny-DEG, a web application that facilitates the exploration and visualization of differentially expressed genes from RNA-seq, was developed. It integrates multi-factor design experiments, allows users to modify the parameters interactively according to experiments purpose and all analysis results can be downloaded directly. Therefore, it serves better for biologists without programming skills. Overall, this project is of great significance to reveal the mechanism of transcriptome differences.

Funding This article was funded by National Natural Science Foundation of China (Grant no. 31800781), China Postdoctoral Science Foundation (Grant no. 2018M631198) and Natural Science Basic Research Program of Shaanxi (Grant no. 2018JQ1012).

Compliance with Ethical Standards

Conflicts of interest The authors declare no competing financial and non-financial interests.

Code availability The source code could be downloaded from github: https://github.com/sufangwang-npu/shiny-DEG.

References

1. Wang GM, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484.RNA-Seq
2. Trapnell et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621
3. Sharad DA, Sztupinszki ZM (2020) Characterization of unique PMEPA1 gene splice variants (isoforms d and e) from RNA Seq profiling provides novel insights into prognostic evaluation of prostate cancer. Oncotarget 11:362–377. https://doi.org/10.18632/oncotarget.27406
4. Marco-puche LS, Benítez J, Trivino JC (2019) RNA-Seq perspectives to improve clinical diagnosis. Front Genet 10:1–7. https://doi.org/10.3389/fgene.2019.01152
5. Meyerson GS, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696. https://doi.org/10.1038/nrg2841
6. Zhang CR, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genom 38:95–109. https://doi.org/10.1016/j.jgg.2011.02.003
7. Koboldt SK, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38. https://doi.org/10.1016/j.cell.2013.09.006
8. Liu H, Xia M, Han L, Zhang Q, Guo Y (2018) GSCALite : a web server for gene set cancer analysis. Bioinformatics 34:3771–3772. https://doi.org/10.1093/bioinformatics/bty411
9. Rau Flister M, Rui H, Auer PL (2019) Exploring drivers of gene expression in The Cancer Genome Atlas. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty551
10. Class HM, Baladandayuthapani V (2017) iDINGO—integrative differential network analysis in genomics with Shiny application. Bioinformatics 34:1243–1245. https://doi.org/10.1093/bioinformatics/btx750
11. Zhang FC, Gan J, Zhu P, Kong L, Li C (2018) iSeq:Web-based RNA-seq data analysis and visualization. Comput Syst Biol Protoc 1754:167–181. https://doi.org/10.1007/978-1-4939-7717-8_10
12. Ren X, Kuan PF (2018) methylGSA : a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty892
13. Khomtchouk HJ, Wahlestedt C (2016) MicroScope : ChIP-seq and RNA-seq software analysis suite for gene expression heatmaps. BMC Bioinformat 17:1–9. https://doi.org/10.1186/s12859-016-1260-x
14. Ernlund SR, Ruggles KV (2018) RIVET : comprehensive graphic user interface for analysis and exploration of genome-wide translatomics data. BMC Bioinformat 19:1–13. https://doi.org/10.1186/s12864-018-5166-z
15. Koeppen SB, Hampton TH (2017) ScanGEO: parallel mining of high-throughput gene expression data. Bioinformatics 33:3500–3501. https://doi.org/10.1093/bioinformatics/btx452
16. Gu Z, Mullighan CG (2018) ShinyCNV: a Shiny/R application to view and annotate DNA copy number variations. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty546
17. Kowalski (2018) shinySISPA : A web tool for defining sample groups using gene sets from multiple-omics data. F1000Research 7:1–11. https://doi.org/10.1268/f1000research.13934.1
18. Nelson SJ, Barnes AP, Minnier J (2017) The START App : a web-based RNAseq analysis and visualization resource. Bioinformatics 33:447–449. https://doi.org/10.1093/bioinformatics/btw624
19. Baker SG, Strachan S, Armstrong M (2018) BLASTmap: a shiny-based application to visualize BLAST results as interactive heat maps and a tool to design gene-specific baits for bespoke target enrichment sequencing. Plant Pathog Fungi Oomycetes Methods Protoc 1848:199–206. https://doi.org/10.1007/978-1-4939-8724-5_14
20. Zagorščak BA, Ramšak Ž, Petek M, Stare T, Gruden K (2018) DiNAR : revealing hidden patterns of plant signalling dynamics using Differential Network Analysis in R. Plant Methods 14:1–9. https://doi.org/10.1186/s13007-018-0345-0
21. Chen (2018) The HTPmod Shiny application enables modeling and visualization of large-scale biological data. Commun Biol 1(1):8. https://doi.org/10.1038/s42003-018-0091-x
22. Mcmurdie PJ, Holmes S (2015) Shiny-phyloseq : Web application for interactive microbiome analysis with provenance tracking. Bioinformatics 31:282–283. https://doi.org/10.1093/bioinformatics/btu616
23. Piccolo WU, Chintapalli SV, Luo C, Shankar K (2018) Dynamic Assessment of Microbial Ecology (DAME): a web app for interactive analysis and visualization of microbial sequencing data. Bioinformatics 34:1050–1052. https://doi.org/10.1093/bioinformatics/btx686
24. Sundararajan KR, Hombach P, Becker M, Schultze JL, Ulas T (2019) "Shiny-Seq : advanced guided transcriptome analysis", BMC Res. Notes 12:1–5. https://doi.org/10.1186/s13104-019-4471-1
25. Ferrero et al (2020) Docker4Circ : a framework for the reproducible characterization of circRNAs from RNA-seq data. Int J Mol Sci 21:1–14. https://doi.org/10.3390/ijms21010293
26. Dillman et al (2013) mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. Nat Neurosci 2:1–9. https://doi.org/10.1038/nn.3332