


SOFTWARE

Open Access

BEAVR: a browser-based tool for the exploration and visualization of RNA-seq data



Pirunthan Perampalam^{1,2} and Frederick A. Dick^{1,3,4*} 

* Correspondence: fdick@uwo.ca

¹London Regional Cancer Program, London Health Sciences Centre, London, ON N6A 5W9, Canada

³Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON N6A 5C1, Canada

Full list of author information is available at the end of the article

Abstract

Background: The use of RNA-sequencing (RNA-seq) in molecular biology research and clinical settings has increased significantly over the past decade. Despite its widespread adoption, there is a lack of simple and interactive tools to analyze and explore RNA-seq data. Many established tools require programming or Unix/Bash knowledge to analyze and visualize results. This requirement presents a significant barrier for many researchers to efficiently analyze and present RNA-seq data.

Results: Here we present **BEAVR**, a Browser-based tool for the Exploration And Visualization of RNA-seq data. **BEAVR** is an easy-to-use tool that facilitates interactive analysis and exploration of RNA-seq data. **BEAVR** is developed in R and uses **DESeq2** as its engine for differential gene expression (DGE) analysis, but assumes users have no prior knowledge of R or **DESeq2**. **BEAVR** allows researchers to easily obtain a table of differentially-expressed genes with statistical testing and then visualize the results in a series of graphs, plots and heatmaps. Users are able to customize many parameters for statistical testing, dealing with variance, clustering methods and pathway analysis to generate high quality figures.

Conclusion: **BEAVR** simplifies analysis for novice users but also streamlines the RNA-seq analysis process for experts by automating several steps. **BEAVR** and its documentation can be found on GitHub at <https://github.com/developerpiru/BEAVR>. **BEAVR** is available as a Docker container at <https://hub.docker.com/r/pirunthan/beavr>.

Keywords: Data visualization, Data exploration, Principle component analysis, Hierarchical gene clustering, Pathway analysis

Background

RNA-sequencing (RNA-seq) has revolutionized molecular biology research in the last decade [1]. RNA-seq is a high-throughput sequencing method that allows for the quantification of gene expression patterns between experimental groups using differential gene expression (DGE) methods [2]. Analysis of DGE may guide the early phases of studies by highlighting transcripts and/or pathways with altered expression in a



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

given experimental system or may be used to assess the downstream impacts of a treatment or other experimental condition. RNA-seq experiments may follow almost any variation of in vitro or in vivo study in which RNA is collected [3]. Most recently, RNA-seq has been employed clinically, including in numerous cancer-related clinical trials [4–6].

Once the wet lab components of an RNA-seq experiment are completed, the data must be analyzed computationally. To date, a multitude of tools are available to researchers depending on the experimental question (e.g. the discovery of novel transcripts or determining gene expression changes) [3, 7]. Regardless of the analysis tool selected, the vast majority of currently available tools require knowledge of programming (C/C++, Perl, Python, R) or shell scripting (Unix/Bash shell). DESeq2, one of the most popular analytical software packages for DGE, is written in R and requires an understanding of this language to manipulate data and visualize results [8]. The requirement for users to navigate one or more computational languages in order to analyze RNA-seq data presents a substantial barrier for many researchers who are adept with respect to the wet lab components of RNA-seq but unfamiliar with the computational aspects.

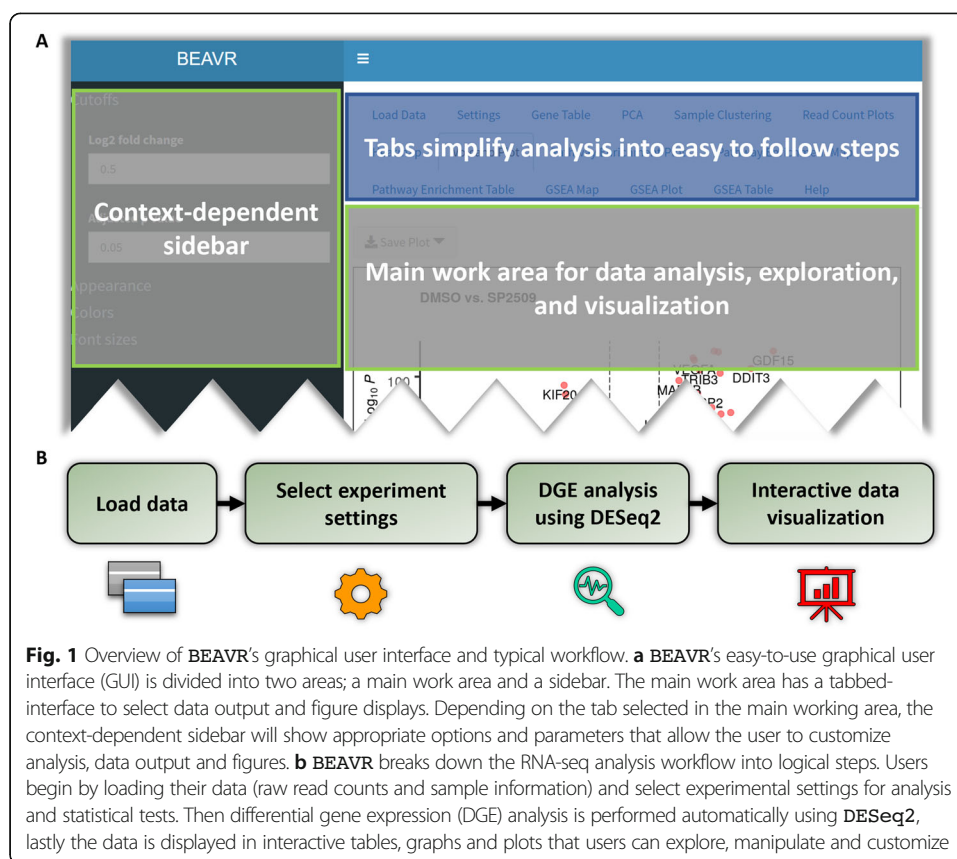
Here, we present BEAVR, a **B**rowser-based tool for the **E**xploration **A**nd **V**isualization of RNA-seq data. BEAVR is an operating system (OS)-independent software package written in R that can run locally on a user's computer or on a remote server. BEAVR provides an easy-to-use graphical frontend to allow both novices and experts to perform DGE analyses on RNA-seq datasets. Specifically, BEAVR simplifies the process of visualization and exploration of results and allows users to generate visually-appealing graphs, tables, plots, heatmaps and pathways maps. At its core, BEAVR uses the heavily-cited DESeq2 as the engine for its analysis. While there is no single superior method for RNA-seq analyses, DESeq2 is an ideal choice because it requires only raw, unnormalized read counts and provides functions to perform DGE and statistical analyses. Our implementation allows for the visualization of PCA plots, read count plots, volcano plots, heatmaps and enriched pathways and facilitates the exploration of DGE results to aid researchers in their study of known gene interactions as well as providing tools for the discovery of novel gene interactions.

Implementation

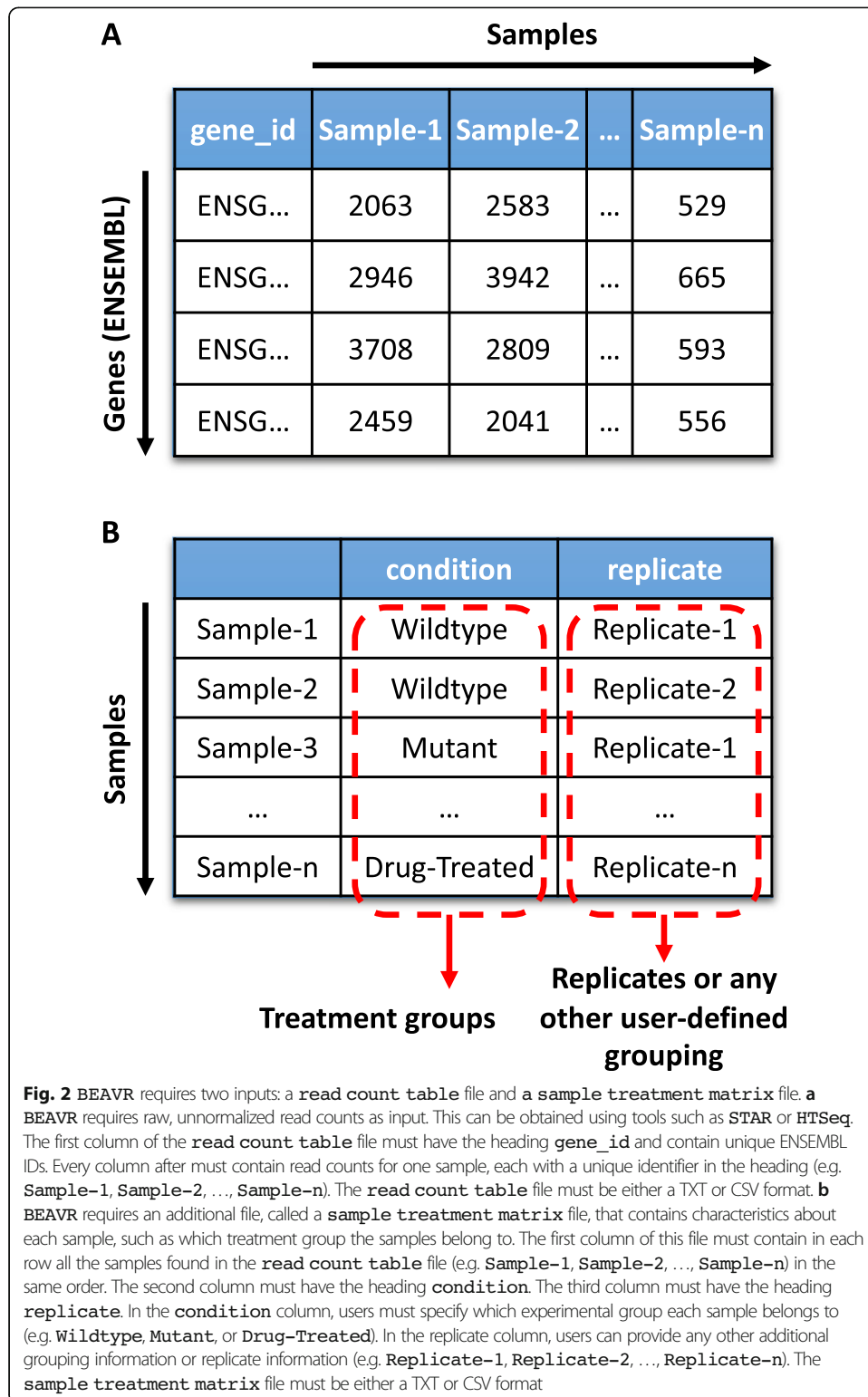
Interface & typical workflow

BEAVR's graphical user interface (GUI) is developed in R using the shiny framework. The layout is divided into a main panel and a sidebar panel (Fig. 1a). The main panel presents the user with a tabbed environment that breaks the workflow of DGE analysis into easy-to-follow logical steps. Depending on which tab is open, the sidebar will display context-dependent parameters that control the output and display of data in the work area of the main panel. The user can manipulate these parameters at any time and the results will be recalculated and updated in real-time, drastically reducing the amount of time required compared to command-line based approaches.

A typical workflow for RNA-seq analysis using BEAVR is shown in Fig. 1b. Briefly, data is loaded into BEAVR, DGE analysis is performed using DESeq2 and the results are visualized in interactive tables, in graphs and other displays. In the Load Data tab,



the user must provide a DESeq2 compatible read count table file containing raw, unnormalized read counts (obtained using alignment tools such as STAR or HTSeq) as well as a sample treatment matrix file (created in a text editor or spreadsheet program). The read count table file (either TXT or CSV) should contain the read quantities for all of the samples in the experiment (Fig. 2a). The first column must contain ENSEMBL identifiers for each gene. The heading for this column must be `gene_id`. The next n columns must contain raw read counts for each of the n samples. The headings for these n columns must be unique sample identifiers (e.g. `wildtype-1`, `wildtype-2`, `wildtype-3`, `mutant-1`, `mutant-2`, `mutant-3`). The sample treatment matrix file (either TXT or CSV) informs BEAVR which columns (samples) in the read count table file belong to which treatment groups (Fig. 2b). This allows multiple replicates to be grouped together across different experimental conditions. The first column must list in each row the sample identifiers for all n columns in the read count table file (e.g. `wildtype-1`, `wildtype-2`, `wildtype-3`, `mutant-1`, `mutant-2`, `mutant-3`). The second column of the sample treatment matrix file specifies which experimental condition each sample belongs to (e.g. `wildtype` and `mutant`, or `untreated` and `drug-treated`). The heading for this column must be `condition`. In the third column, the user may specify any additional characteristics for each sample, such as replicate numbers/letters or genotype groups (e.g. `replicate-A`, `replicate-B`, `replicate-C`). The heading for this column must be `replicate`. Both the read count table file



and the **sample treatment matrix** file must contain at least two experimental conditions with a minimum of 2 samples each. Treatment groups do not need to contain the same number of samples in each group.

In the **Settings** tab, the user must select a control condition and a treatment condition (condition choices are loaded from those available in the `sample treatment matrix` file). For DGE analyses, DESeq2 is used to compare the selected treatment condition against the selected control condition. The user may specify a minimum cut-off for reads if desired (reads below this cutoff value are dropped before analysis), specify a false discovery rate (FDR) to determine adjusted p values (p_{adj}) and also specify an effect size shrinkage method using DESeq2 [8] or `apeglm` (approximate posterior estimation) [9].

Representation of results & data exploration

Clicking on the **Gene Table** tab will initiate automated DGE analysis using the parameters specified by the user. A progress bar will be shown in the bottom right of the main work area. Upon completion, an interactive table displays the results including gene IDs as HUGO Gene Nomenclature Committee (HGNC) symbols, \log_2 fold changes (LFC), p values and p_{adj} values for each gene. Controls in the sidebar may be used to filter the table as desired and a copy may be saved using the **Download Table** button.

Visualization of all plots is implemented using `ggplot2`. The **PCA** tab will generate a principle component analysis (PCA) plot and display all the samples found in the `read count table` file. In the **Sample Clustering** tab, the user can select a distance measurement method to use (Pearson correlation, Euclidean, Maximum, Manhattan, Canberra, Binary, or Minkowski) which will compute a distance matrix using the `ComplexHeatmap` and `dist` packages and display the sample variation as a heatmap. The **Read Count Plots** tab will generate normalized read count plots, either as boxplots or jitter plots, for desired genes. The user can enter gene names separated by a comma and change the grid layout as desired (use a 1×1 grid for a single plot or increase the grid size as necessary to fit multiple plots). The **Heatmap** tab will allow the user to generate a heatmap with gene clustering for the top n significantly variable genes (where n is a user-defined number), or for any list of genes entered by the user. Dependence of the variance on the mean is removed using either variance stabilization (`vst`) or regularized logarithm (`rlog`) transformations [8] as specified by the user. The user can also specify a hierarchical clustering method (Ward.D/D2, Single, Complete, Average, McQuitty, Median, or Centroid) to be used by the `hclust` package (for row and/or column clustering) and a distance measurement method as described above. The **Volcano Plot** tab will generate a volcano plot using the `EnhancedVolcano` package to illustrate differentially-expressed genes that meet the user-defined LFC and p_{adj} cutoffs for the control and treatment conditions specified on the **Settings** tab. Pathway over-representation analysis and gene set enrichment analysis (GSEA) are performed using the `ReactomePA` and `enrichplot` packages [10] and figures are shown in the **Pathway Enrichment Plot**, **Pathway Enrichment Map**, **GSEA Plot** and **GSEA Map** tabs with the tabular results being displayed in the **Pathway Enrichment Table** and **GSEA Table** tabs. All customization options are presented in the sidebar and allow users to control many parameters when plotting figures, including the ability to customize colors, font sizes and legend positions and directions (horizontal or vertical) for all figures. The size and aspect ratio of all figures can be adjusted by clicking

and dragging the outside edges of the plot area. The `Save Plot` button located above every plot allows figures to be saved in multiple formats (JPEG, PDF, PNG, SVG, TIFF) while the `Download Table` button in the sidebar allows data from any table to be saved (CSV).

Installation

Since BEAVR is developed in R (+ 3.5), it is OS-independent and runs on Linux, Mac OS and Windows. We provide several methods to install and use BEAVR depending on user preference: 1) the easiest method for those unfamiliar with R is to install Docker (<https://docker.com>) and use our Docker container (<https://hub.docker.com/r/pir-unthan/beavr>) which comes pre-installed with all of the required components; or 2) users can use our OS-specific scripts to install and configure R with all of the required packages for BEAVR; or 3) users who already have R installed can download BEAVR from GitHub. Additionally, system administrators may install BEAVR in a multi-user server environment which is useful for research groups that want to have a centralized server for BEAVR. This is implemented using ShinyProxy (<https://shinyproxy.io>) and Docker which provide a secure, sandboxed environment for every connected user. We provide automated install scripts on GitHub to easily accomplish this and system administrators can customize the installation to their specific network requirements. Each of these methods simplify and streamline setup for novice and expert users alike and are well-documented on the GitHub page for BEAVR located at <https://github.com/developerpiru/BEAVR>.

Run time consideration

Computation time is dependent on the user's device specifications since all DGE analyses, statistical tests and visualization steps are performed locally (or the server specifications when running BEAVR on a shared server). For a typical mammalian RNA-seq experiment containing two experimental groups with three replicates each using the human genome as a reference (88 million reads total), automated calculations will take approximately 1 min with a dual-core Intel Core i5 CPU and 4 GB RAM or approximately 30 s with a 6-core Core i7 and 16 GB RAM. Generation of each figure, as well as subsequent modifications thereto, will take a few additional seconds. These short processing times will allow users to repeatedly manipulate experimental settings to recalculate DGE as desired with different parameters. Users may then explore the results, generating figures and filtering and downloading the data for downstream applications.

Results & discussion

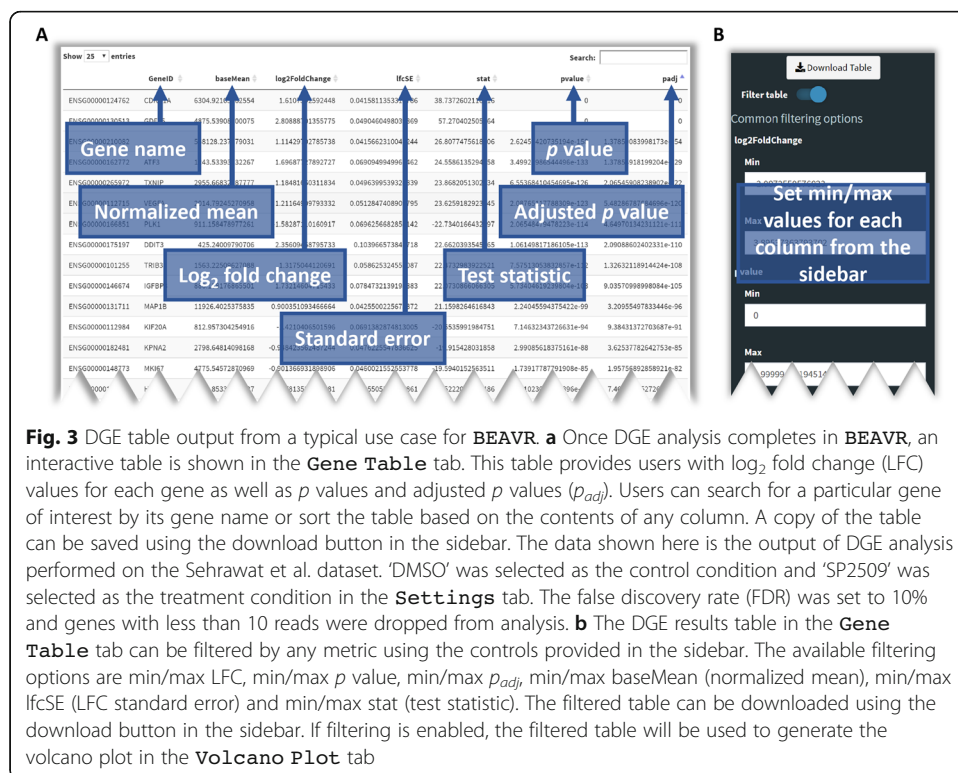
A typical use case

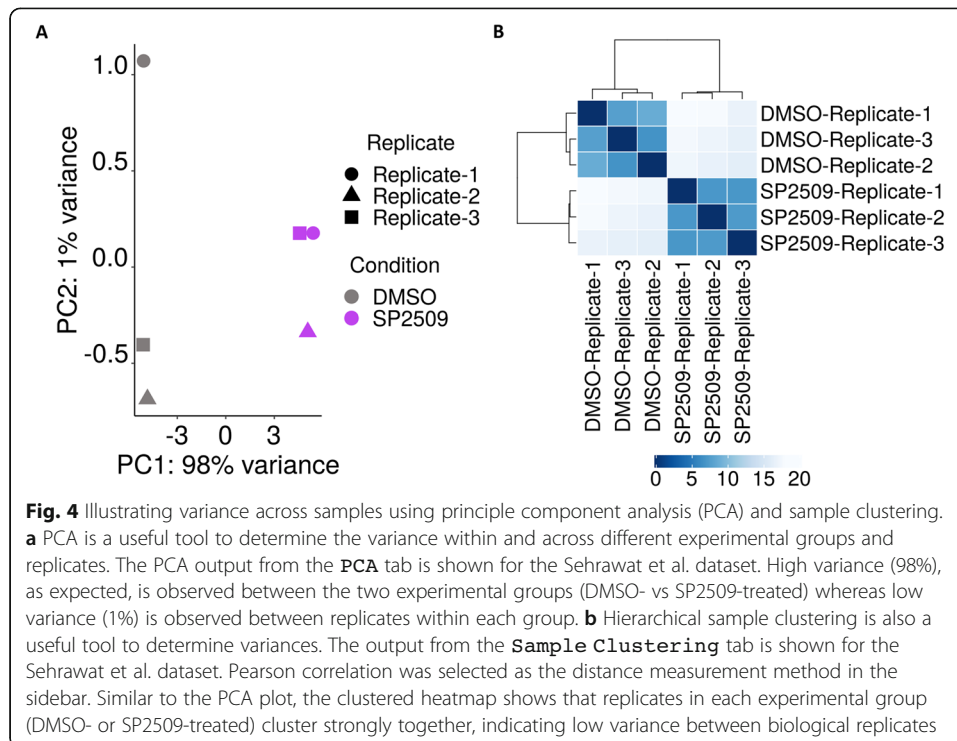
To demonstrate a typical use case for BEAVR, we utilized a previously published RNA-seq dataset by Sehwat et al. [11]. In this study, LNCaP cell cultures were treated with either DMSO or SP2509 (a small molecule lysine-specific demethylase 1 [LSD1] inhibitor) for 24 h [11]. RNA-seq was performed on RNA harvested from triplicate cell cultures corresponding to each treatment condition. We downloaded raw, unnormalized read counts from GEO (GSE59009) and merged the read counts from all samples to make a single read count table file (TXT). We created a sample treatment

matrix file (CSV) using Microsoft Excel to specify the treatment condition group (either DMSO or SP2509) and replicate number for each sample. Once these two files were prepared, they were loaded into BEAVR from the Load data tab. In the Settings tab, we selected ‘DMSO’ as the control condition and ‘SP2509’ as the treatment condition. The FDR was set to 10% and the minimum threshold to drop reads was set to 10.

Figure 3a shows the DGE results from the Gene table tab, which has been sorted by ascending p_{adj} values. This table can be saved as-is or it can be filtered. For example, it is often desirable to have a list of only those genes that exceed a specific LFC threshold (e.g. ± 1.0) and fall below a p_{adj} threshold (e.g. < 0.05). These values can be set using the sidebar (Fig. 3b) and the results table will be updated automatically to display genes meeting the selected criteria. These parameters also instruct the thresholds used in generating the volcano plot and pathway analyses.

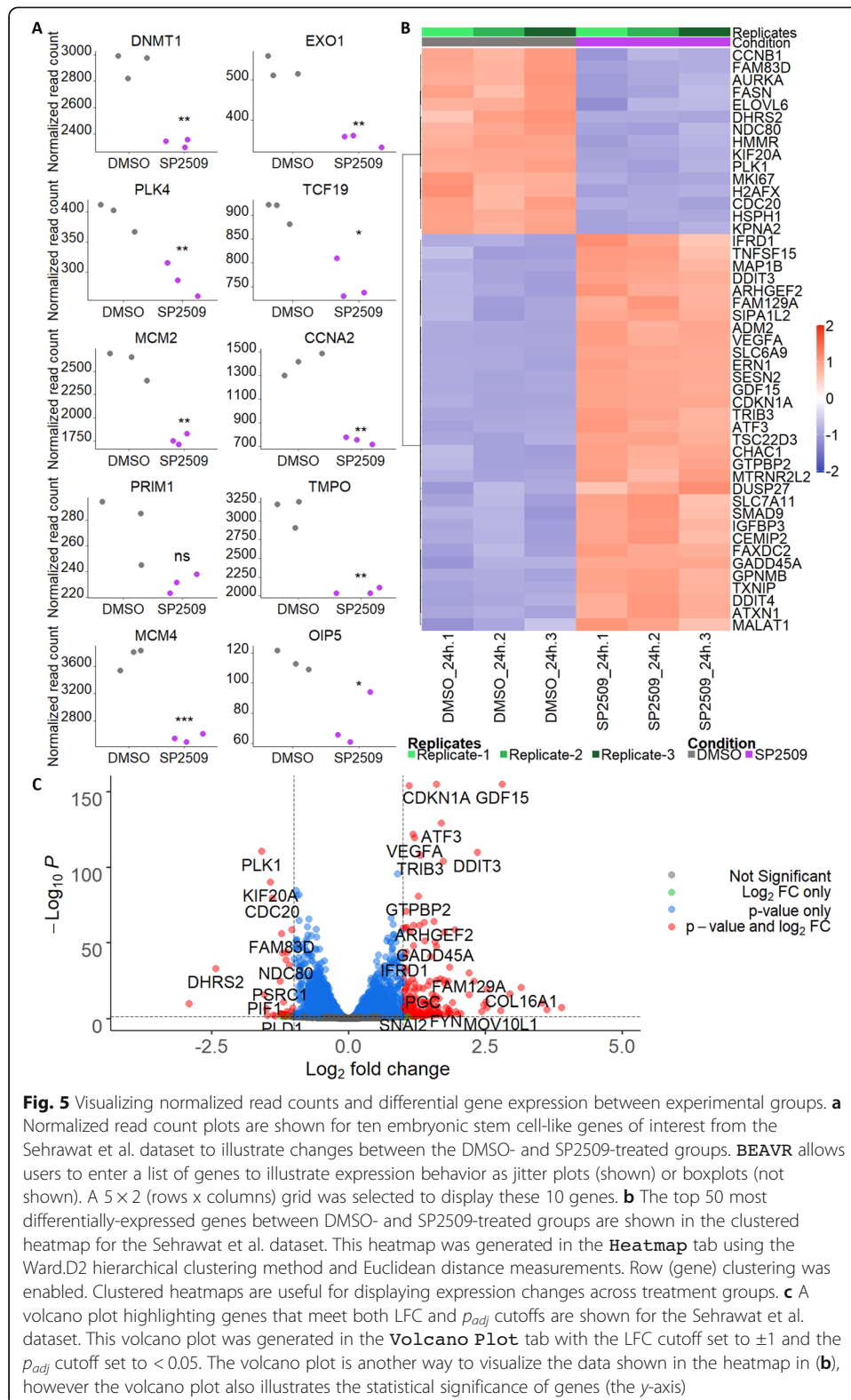
PCA is an important consideration in RNA-seq analysis for small and large studies. Depending on the experimental design, PCA plots can be used for quality control or as a discovery tool [12]. In studies with only two control groups and just two or three biological replicates, it can inform researchers of replicates that are not congruent and have high variance which can skew results and reduce statistical power. In larger studies, it can provide insight into the heterogeneity within experimental conditions. The PCA Plot tab displays a PCA plot from our example dataset. The plot shows that there is a very small amount of variance (1%) between replicates within each experimental group (DMSO- or SP2509-treated), while there is very large variance, as expected, between the two experimental groups (98%) (Fig. 4a). Further quality control and insight into sample and replicate variation can be interrogated through a distance





matrix and subsequent sample clustering. We defined the parameters in the **Sample Clustering** tab to compute Pearson correlation distances and the result is shown in Fig. 4b. Replicates in the same experimental group cluster together and are very similar to each other, indicating very low variance. Together, these two graphs provide researchers with useful information about experimental groups and consistency of biological replicates.

Sehwat et al. found inhibition of LSD1 in LNCaP cells caused downregulation of previously characterized embryonic stem cell-like genes [11, 13]. Using the **Read Count Plots** tab, we explored the normalized read counts of these genes and generated plots that showed reduced normalized reads in the SP2509-treated cells compared to DMSO-treated cells (Fig. 5a). In situations where genes or pathways of interest are already known, read count plots can be used as a tool to investigate changes in gene expression across samples. However, RNA-seq is also used in experimental systems to inform researchers of genes and pathways that may be of interest. For such purposes, a heatmap with gene clustering or a volcano plot are useful tools. The **Heatmap** tab generates heatmaps for the top n genes (where n is a user-defined number) or for specific genes entered by the user. Figure 5b shows the top 50 most differentially-expressed genes after variance stabilization with hierarchical clustering performed across rows (Ward.D2 method). This provides information on the most strongly upregulated and downregulated genes. Although the data for a heatmap is transformed and variance is stabilized, it does not provide information on significance (p values or p_{adj}) [8]. The volcano plot from the **Volcano Plot** tab illustrates genes that meet a specified LFC threshold as well as a p_{adj} threshold (Fig. 5c). We set the LFC threshold to ± 1.0 and the p_{adj} cutoff to < 0.05 . Genes highlighted in red (meeting both the LFC and p_{adj} cutoffs) were also



found in the heatmap, demonstrating the usefulness of heatmaps and volcano plots and how the two can be used together for discovery of novel gene expression patterns.

Following identification of upregulated and downregulated genes, it is useful to perform pathway enrichment or gene set enrichment analysis (GSEA) [14, 15] to identify important pathways of interest that will inform investigators of downstream experiments. The **Pathway Enrichment Plot** tab performs over-representation analysis and produces either a dot plot or bar graph of the top n pathways (where n is a user-defined number) (Fig. 6a). The **Pathway Enrichment Map** tab provides a broader look at all enriched pathways using an interconnected network map (Fig. 6b) that shows the results of over-representation analysis, however users may also wish to perform GSEA on the **GSEA Map** tab. The **GSEA Plot** tab displays a plot of the running enrichment score for a specific enriched pathway as defined by the user (Fig. 6c). The input data used to generate these figures is the filtered or unfiltered data from the **Gene Table** tab (we filtered the data using $LFC < 0$ and $p_{adj} < 0.05$). The pathways identified in Fig. 6a-c are consistent with the most downregulated genes shown in the heatmap (Fig. 5b) and volcano plot (Fig. 5c) (such as *H2AX*, *CDC20*, *CCNBL*, *AURKA*) and indicate the most significantly enriched pathways among downregulated genes are related to cell cycle and DNA replication processes. Together, the read count plots, heatmap, volcano plot and pathway plots inform researchers of gene expression changes and provide insight into which genes and pathways may play an important role in their experimental system.

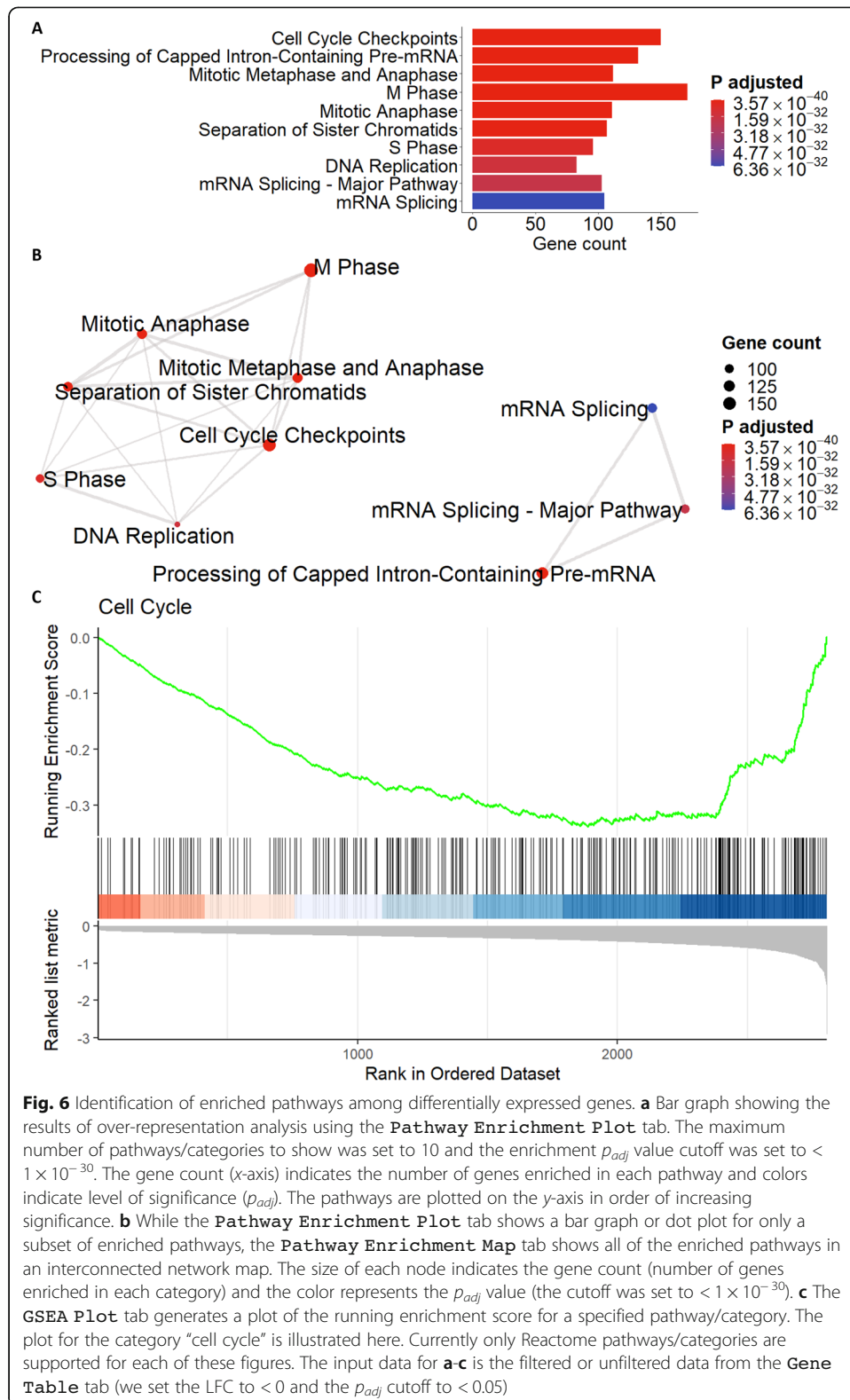
Future work

DGE analyses computes differences between two groups at a time, such as **Wildtype** and **Single-knockout**, even though users can load data files containing > 2 groups (e.g. **Wildtype**, **Single-knockout** and **Double-knockout**). Currently, users must perform one comparison first (e.g. **Wildtype** vs **Single-knockout**), download the results and then perform another comparison (e.g. **Wildtype** vs **Double-knockout**) and download the new results. Users must then manually perform comparisons outside of BEAVR to identify overlapping or non-overlapping genes. Future updates to BEAVR will allow users to perform multiple DGE analyses and allow them to interact with both results at once to perform direct comparisons within BEAVR. Implementation of additional plotting tools, such as Euler or Venn diagrams, will allow for the visualization of overlapping or non-overlapping dysregulated genes across different comparisons such as **Wildtype** vs **Single-knockout** and **Wildtype** vs **Double-knockout**. These overlapping or non-overlapping datasets can then be used to perform pathway analysis or GSEA within BEAVR.

Presently, BEAVR only supports Reactome categories for pathway analysis and GSEA. Future updates will enable support for Gene Ontology (GO) [16], Disease Ontology (DO) [17], KEGG [18], WikiPathways [19] and Molecular Signature Database (MSigDb) [14, 20] to provide users with more options.

Conclusions

RNA-seq analyses has largely relied on command-line-driven tools, such as DESeq2 [8], EdgeR [21] or ALDEx [22], thereby creating a barrier to entry for scientists wishing to conduct RNA-seq analyses. Here we presented BEAVR, a graphically-driven tool that greatly simplifies DGE analyses through a logical workflow that makes use of DESeq2



as the core DGE engine. BEAVR is easy-to-use and allows researchers to not only quickly and easily change experimental parameters in real-time to visualize results, but also provides an intuitive interface for researchers to explore their results in-depth and

generate highly customizable figures. Various other tools have been developed to provide users with graphical interfaces for RNA-seq analyses, most notably GENAVi [23], START [24], iDEP [25], DEBrowser [26], DEIVA [27] and DEApp [28]. While these tools have undoubtedly provided a significant evolution in RNA-seq analysis tools, we found that BEAVR offers meaningful advantages in comparison. Specifically, the ease of installation and usage, combined with more flexibility in data output features are important advancements. None of these programs offers each of our key features in one complete package, such as filtering capabilities of gene lists, all of the different data displays that BEAVR provides (heat-map, PCA plots, etc.), the ability to customize and export figures in as many formats, or the ability to integrate pathway analysis. Based on these differences we expect BEAVR will be widely utilized.

BEAVR was developed to be simple enough for novices, yet fast and powerful enough for experts to streamline and automate DGE analyses. Even with modest computing power by today's standards, BEAVR is capable of completing analyses within minutes, allowing researchers to quickly automate analyses of large datasets. With uses for RNA-seq continuing to expand — both experimentally and clinically — BEAVR is well-positioned to allow analysis of these datasets to be quick and efficient, while providing the latitude for customization as per the user's requirements.

Availability and requirements

Project name: BEAVR

Project home page: <https://github.com/developerpiru/BEAVR> and <https://hub.docker.com/r/pirunthan/beavr>

Project documentation: <https://github.com/developerpiru/BEAVR/blob/master/README.md>

Operating system: Linux, Mac OS, Windows

Programming language: R

Other requirements: R 3.5 or higher, web browser

License: GNU General Public License v3.0

Any restrictions to use by non-academics: None

Abbreviations

BEAVR: Browser-based exploration and visualization of RNA-seq data; DGE: Differential gene expression; DO: Disease Ontology; FDR: False discovery rate; GO: Gene Ontology; GSEA: Gene set enrichment analysis; GUI: Graphical user interface; HGNC: HUGO Gene Nomenclature Committee; KEGG: Kyoto Encyclopedia of Genes and Genomes; LFC: Log₂ fold change; LSD1: Lysine-specific demethylase 1; MSigDb: Molecular Signature Database; OS: Operating system; p_{adj} : Adjusted p value; PCA: Principle component analysis; RNA-seq: RNA-sequencing

Acknowledgements

We wish to thank Dr. John W. Barrett, Dr. Haider M. Hassan and Michael V. Roes (London Regional Cancer Program, London Health Sciences Centre) for beta testing BEAVR and providing key feedback.

Authors' contributions

PP wrote the software, tested it with sample data, and co-wrote the manuscript. FAD assisted in data analysis and display features, and co-wrote the manuscript. All authors approved the final version of the manuscript and are available to answer questions about its content.

Funding

PP was supported by CaRTT. FAD is the Wolfe Senior Fellow in Tumor Suppressor Genes at Western University. This study was funded by the Canadian Institutes of Health Research and funds provided by the London 'Run for Ovarian Cancer'. No funding sources participated in the design, data collection, analysis, interpretation of data, or preparation of the manuscript.

Availability of data and materials

The dataset used in this article is available in the GEO repository, (GSE59009).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to declare.

Author details

¹London Regional Cancer Program, London Health Sciences Centre, London, ON N6A 5W9, Canada. ²Department of Biochemistry, University of Western Ontario, London, ON N6A 5C1, Canada. ³Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON N6A 5C1, Canada. ⁴Children's Health Research Institute, London, ON N6A 4V2, Canada.

Received: 28 January 2020 Accepted: 18 May 2020

Published online: 29 May 2020

References

1. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2017;8(1):10.
2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*. 2019;20(11):631–56.
3. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17:13.
4. Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, Romano A. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18(2):308.
5. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, Sekacheva M, Kim E, Gaifullin N, Giese A, et al. RNA sequencing for research and diagnostics in clinical oncology. *Semin Cancer Biol*. 2019;60:311–23.
6. Marco-Puche G, Lois S, Benitez J, Trivino JC. RNA-Seq perspectives to improve clinical diagnosis. *Front Genet*. 2019;10:1152.
7. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS One*. 2017;12(12):e0190152.
8. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
9. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 2019;35(12):2084–92.
10. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol BioSyst*. 2016;12(2):477–9.
11. Sehrawat A, Gao L, Wang Y, Bankhead A 3rd, McWeeney SK, King CJ, Schwartzman J, Urrutia J, Bisson WH, Coleman DJ, et al. LSD1 activates a lethal prostate cancer gene network independently of its demethylase function. *Proc Natl Acad Sci U S A*. 2018;115(18):E4179–88.
12. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. *Brief Bioinform*. 2011;12(6):714–22.
13. Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, Chang HY. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*. 2008;2(4):333–44.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;102(43):15545.
15. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267–73.
16. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2018;47(D1):D330–8.
17. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40(Database issue):D940–6.
18. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47(D1):D590–5.
19. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*. 2018;46(D1):D661–7.
20. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417–25.
21. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*. 2010;26(1):139–40.
22. Fernandes AD, MacKlaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One*. 2013;8(7):e67019.
23. Reyes ALP, Silva TC, Coetzee SG, Plummer JT, Davis BD, Chen S, Hazelett DJ, Lawrenson K, Berman BP, Gayther SA, et al. GENAVI: a shiny web application for gene expression normalization, analysis and visualization. *BMC Genomics*. 2019;20(1):745.
24. Nelson JW, Sklenar J, Barnes AP, Minnier J. The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics (Oxford, England)*. 2017;33(3):447–9.
25. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*. 2018;19(1):534.

26. Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics*. 2019;20(1):6.
27. Harshbarger J, Kratz A, Carninci P. DEIVA: a web application for interactive visual analysis of differential gene expression profiles. *BMC Genomics*. 2017;18(1):47.
28. Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol Med*. 2017;12(1):2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

