5-29-2020

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

Pirunthan Perampalam
London Regional Cancer Program

Frederick A. Dick
London Regional Cancer Program, fdick@uwo.ca

Perampalam and Dick *BMC Bioinformatics* (2020) 21:221
https://doi.org/10.1186/s12859-020-03549-8

BMC Bioinformatics

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
[1]London Regional Cancer Program, London Health Sciences Centre, London, ON N6A 5W9, Canada
[3]Department 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 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 **R**NA-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,

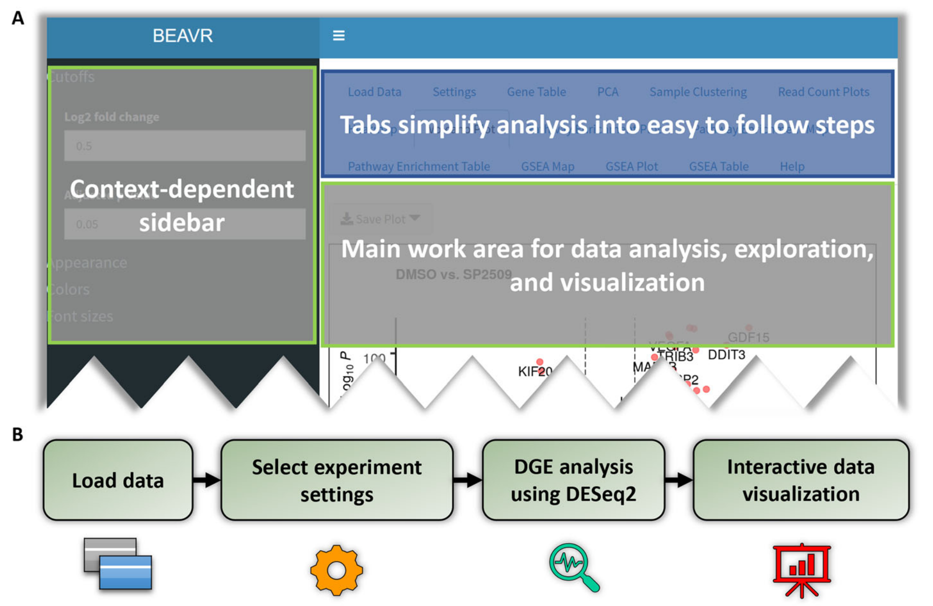
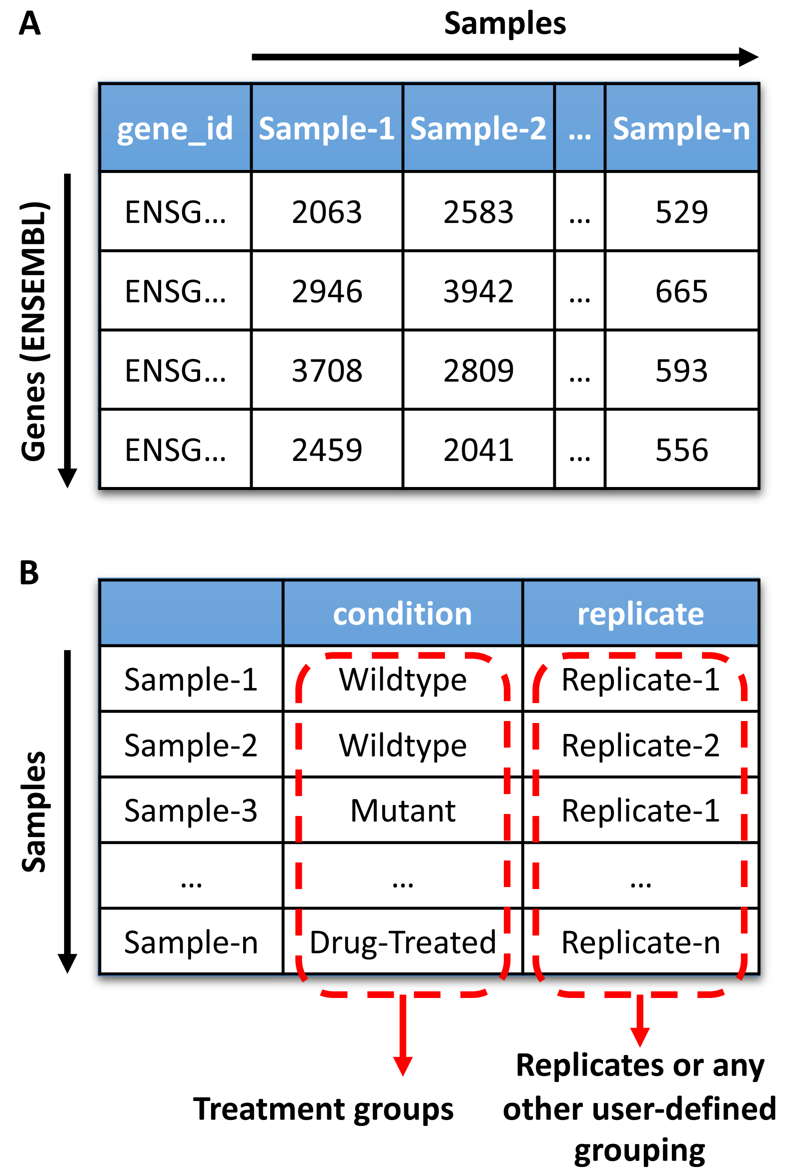


Fig. 1 Overview of BEAVR's graphical user interface and typical workflow. **a** BEAVR's easy-to-use graphical user interface (GUI) is divided into two areas; a main work area and a sidebar. The main work area has a tabbed-interface to select data output and figure displays. Depending on the tab selected in the main working area, the context-dependent sidebar will show appropriate options and parameters that allow the user to customize analysis, data output and figures. **b** BEAVR breaks down the RNA-seq analysis workflow into logical steps. Users begin by loading their data (raw read counts and sample information) and select experimental settings for analysis and statistical tests. Then differential gene expression (DGE) analysis is performed automatically using DESeq2, lastly the data is displayed in interactive tables, graphs and plots that users can explore, manipulate and customize

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



gene_id	Sample-1	Sample-2	…	Sample-n
ENSG…	2063	2583	…	529
ENSG…	2946	3942	…	665
ENSG…	3708	2809	…	593
ENSG…	2459	2041	…	556

	condition	replicate
Sample-1	Wildtype	Replicate-1
Sample-2	Wildtype	Replicate-2
Sample-3	Mutant	Replicate-1
…	…	…
Sample-n	Drug-Treated	Replicate-n

Fig. 2 `BEAVR` requires two inputs: a **`read count table`** file and **a `sample treatment matrix`** file. **a** `BEAVR` requires raw, unnormalized read counts as input. This can be obtained using tools such as `STAR` or `HTSeq`. The first column of the **`read count table`** file must have the heading `gene_id` and contain unique ENSEMBL IDs. Every column after must contain read counts for one sample, each with a unique identifier in the heading (e.g. `Sample-1`, `Sample-2`, …, `Sample-n`). The **`read count table`** file must be either a TXT or CSV format. **b** `BEAVR` requires an additional file, called a **`sample treatment matrix`** file, that contains characteristics about each sample, such as which treatment group the samples belong to. The first column of this file must contain in each row all the samples found in the **`read count table`** file (e.g. `Sample-1`, `Sample-2`, …, `Sample-n`) in the same order. The second column must have the heading `condition`. The third column must have the heading `replicate`. In the `condition` column, users must specify which experimental group each sample belongs to (e.g. `Wildtype`, `Mutant`, or `Drug-Treated`). In the replicate column, users can provide any other additional grouping information or replicate information (e.g. `Replicate-1`, `Replicate-2`, …, `Replicate-n`). The `sample treatment matrix` file must be either a TXT or CSV format

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 cutoff 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/pirunthan/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 re-calculate DGE as desired with different parameters. Users may then explore the results, generating figures and filtering and downloading the data for downstream applications.

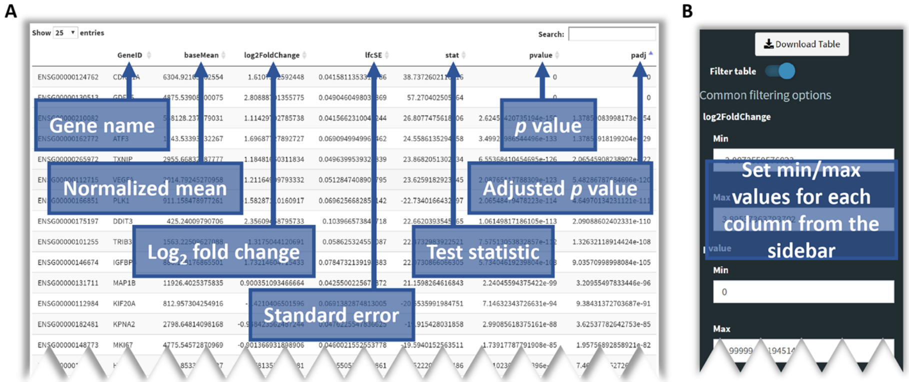
Results & discussion

A typical use case

To demonstrate a typical use case for `BEAVR`, we utilized a previously published RNA-seq dataset by Sehrawat 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



	GeneID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000124762	CD[illegible]A	6304.921[illegible]2554	1.610[illegible]2592448	0.04158113533[illegible]86	38.7372602[illegible]16	0	0
ENSG00000130513	GDF[illegible]	4875.53908[illegible]00075	2.808887[illegible]1355775	0.049046049803[illegible]869	57.270402505[illegible]64	0	0
ENS[illegible]000210082	[illegible]	5[illegible]8128.237[illegible]79031	1.114297[illegible]2785738	0.041566231004[illegible]244	26.8077475618[illegible]06	2.6245[illegible]420735194e-15[illegible]	1.3785[illegible]083998173e-[illegible]54
ENSG00000162772	ATF3	1[illegible]43.53393[illegible]32267	1.696877[illegible]7892727	0.069094994996[illegible]462	24.5586135294[illegible]58	3.4992[illegible]986544496e-133	1.3785[illegible]918199204e-[illegible]29
ENSG00000265972	TXNIP	2955.66832[illegible]87777	1.184816[illegible]0311834	0.049639953932[illegible]839	23.8682051302[illegible]34	6.55368410454695e-126	2.06545908238902e-[illegible]22
EN[illegible]G00000112715	VEGFA	2014.79245270958	1.211649[illegible]9793332	0.051284740890[illegible]795	23.6259182923[illegible]45	2.08765117788309e-123	5.48286787684696e-120
EN[illegible]G00000166851	PLK1	911.158478977261	-1.582871[illegible]0160917	0.069625668285[illegible]142	-22.7340166432[illegible]97	2.06548479478223e-114	4.64970134231121e-111
ENSG00000175197	DDIT3	425.24009790706	2.356094[illegible]8795733	0.10396657384[illegible]718	22.6620393545[illegible]65	1.06149817186105e-113	2.09088602402331e-110
ENSG00000101255	TRIB3	1563.22509627088	1.3175044120691	0.05862532455[illegible]087	22.[illegible]732983922521	7.57513053832857e-11[illegible]	1.32632118914424e-108
ENSG00000146674	IGFBP[illegible]	86[illegible]3176865501	1.73214604[illegible]15433	0.078473213919[illegible]883	22.[illegible]730866066305	5.73404619239804e-10[illegible]	9.03570998998084e-105
ENSG00000131711	MAP1B	11926.4025375835	0.900351093466664	0.042550022567[illegible]872	21.1598264616843	2.24045594375422e-99	3.20955497833446e-96
ENSG00000112984	KIF20A	812.957304254916	-[illegible].4210406501596	0.0691382874813005	-20.[illegible]535991984751	7.14632343726631e-94	9.38431372703687e-91
ENSG00000182481	KPNA2	2798.64814098168	-0.9[illegible]8423562487244	0.0476225547836625	-1[illegible].915428031858	2.99085618375161e-88	3.62537782642753e-85
ENSG00000148773	MKI67	4775.54572870969	-0.901366931898906	0.0460021552553778	-19.5940152563511	1.73917787791908e-85	1.95756892858921e-82
[illegible]0000[illegible]	H[illegible]	[illegible]8533[illegible]7	[illegible]8135[illegible]81	[illegible]5505[illegible]861	[illegible]5222[illegible]86	[illegible]1023[illegible]396e-[illegible]	7.4[illegible]5272[illegible]

Filter table ●

Common filtering options

log2FoldChange

Min

[illegible]

Max

3.89537363793702

p value

Min

0

Max

[illegible]99999[illegible]19451[illegible]

Fig. 3 DGE table output from a typical use case for **`BEAVR`**. **a** Once DGE analysis completes in **`BEAVR`**, an interactive table is shown in the **`Gene Table`** tab. This table provides users with $\log_2$ fold change (LFC) values for each gene as well as *p* values and adjusted *p* values (p_{adj}). Users can search for a particular gene of interest by its gene name or sort the table based on the contents of any column. A copy of the table can be saved using the download button in the sidebar. The data shown here is the output of DGE analysis performed on the Sehrawat et al. dataset. 'DMSO' was selected as the control condition and 'SP2509' was selected as the treatment condition in the **`Settings`** tab. The false discovery rate (FDR) was set to 10% and genes with less than 10 reads were dropped from analysis. **b** The DGE results table in the **`Gene Table`** tab can be filtered by any metric using the controls provided in the sidebar. The available filtering options are min/max LFC, min/max *p* value, min/max p_{adj}, min/max baseMean (normalized mean), min/max lfcSE (LFC standard error) and min/max stat (test statistic). The filtered table can be downloaded using the download button in the sidebar. If filtering is enabled, the filtered table will be used to generate the volcano plot in the **`Volcano Plot`** tab

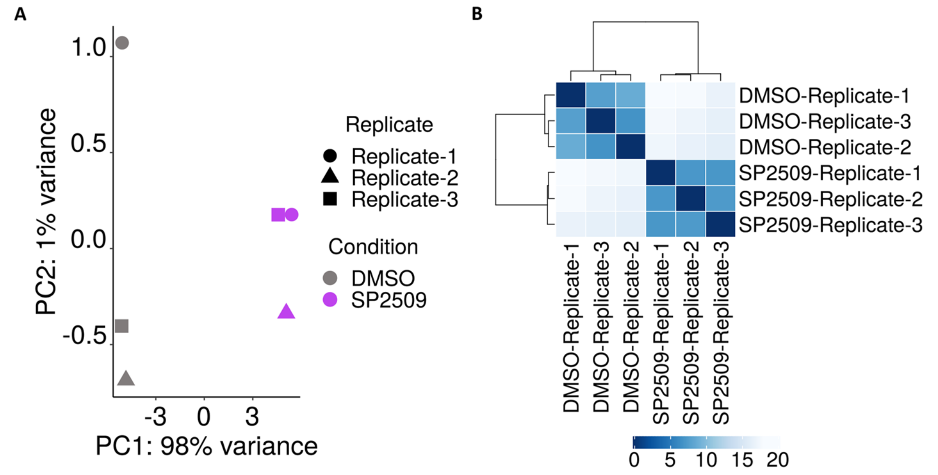


Fig. 4 Illustrating variance across samples using principle component analysis (PCA) and sample clustering. **a** PCA is a useful tool to determine the variance within and across different experimental groups and replicates. The PCA output from the `PCA` tab is shown for the Sehrawat et al. dataset. High variance (98%), as expected, is observed between the two experimental groups (DMSO- vs SP2509-treated) whereas low variance (1%) is observed between replicates within each group. **b** Hierarchical sample clustering is also a useful tool to determine variances. The output from the `Sample Clustering` tab is shown for the Sehrawat et al. dataset. Pearson correlation was selected as the distance measurement method in the sidebar. Similar to the PCA plot, the clustered heatmap shows that replicates in each experimental group (DMSO- or SP2509-treated) cluster strongly together, indicating low variance between biological replicates

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.

Sehrawat 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

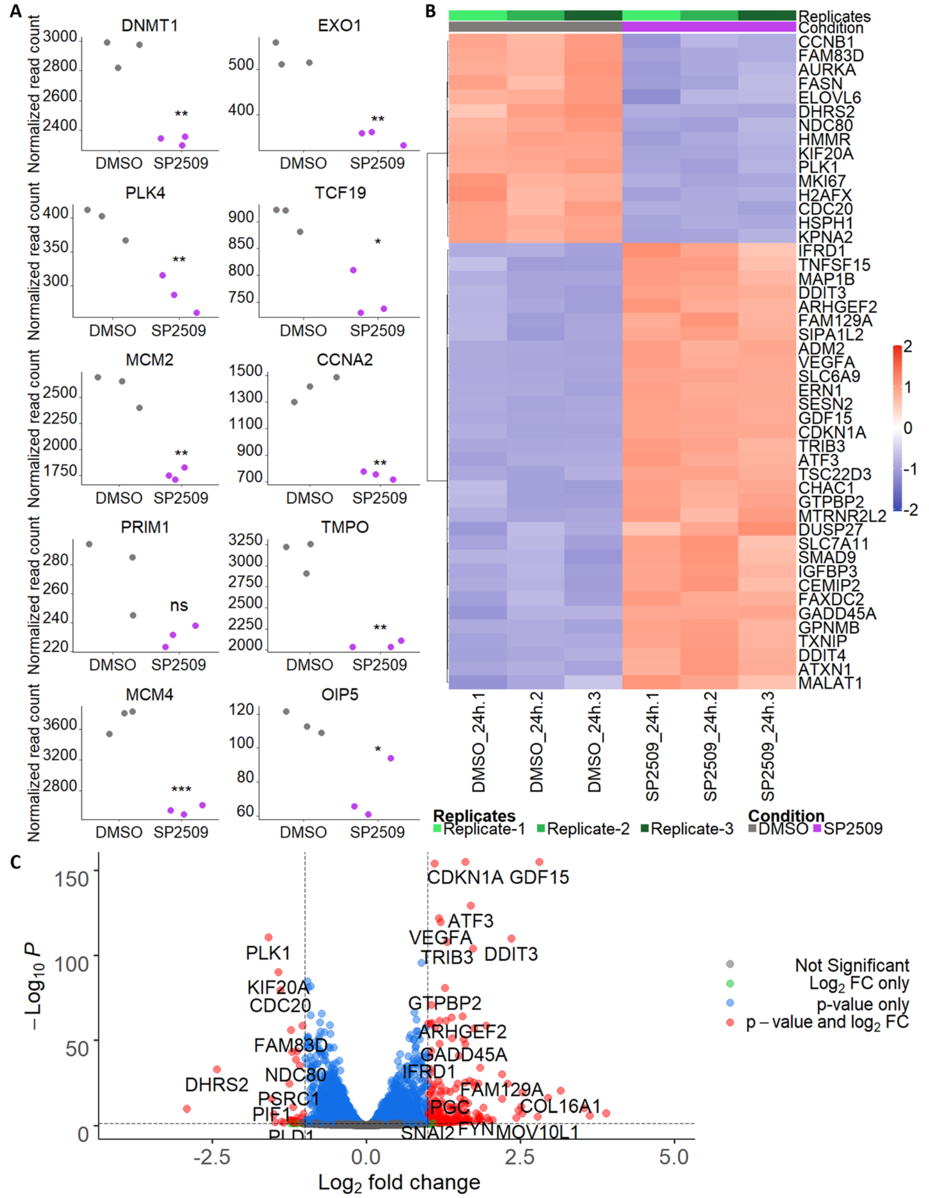


Fig. 5 Visualizing normalized read counts and differential gene expression between experimental groups. **a** Normalized read count plots are shown for ten embryonic stem cell-like genes of interest from the Sehrawat et al. dataset to illustrate changes between the DMSO- and SP2509-treated groups. `BEAVR` allows users to enter a list of genes to illustrate expression behavior as jitter plots (shown) or boxplots (not shown). A 5 × 2 (rows x columns) grid was selected to display these 10 genes. **b** The top 50 most differentially-expressed genes between DMSO- and SP2509-treated groups are shown in the clustered heatmap for the Sehrawat et al. dataset. This heatmap was generated in the `Heatmap` tab using the Ward.D2 hierarchical clustering method and Euclidean distance measurements. Row (gene) clustering was enabled. Clustered heatmaps are useful for displaying expression changes across treatment groups. **c** A volcano plot highlighting genes that meet both LFC and p_{adj} cutoffs are shown for the Sehrawat et al. dataset. This volcano plot was generated in the `Volcano Plot` tab with the LFC cutoff set to ±1 and the p_{adj} cutoff set to < 0.05. The volcano plot is another way to visualize the data shown in the heatmap in (**b**), however the volcano plot also illustrates the statistical significance of genes (the *y*-axis)

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, CCNB1, 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`

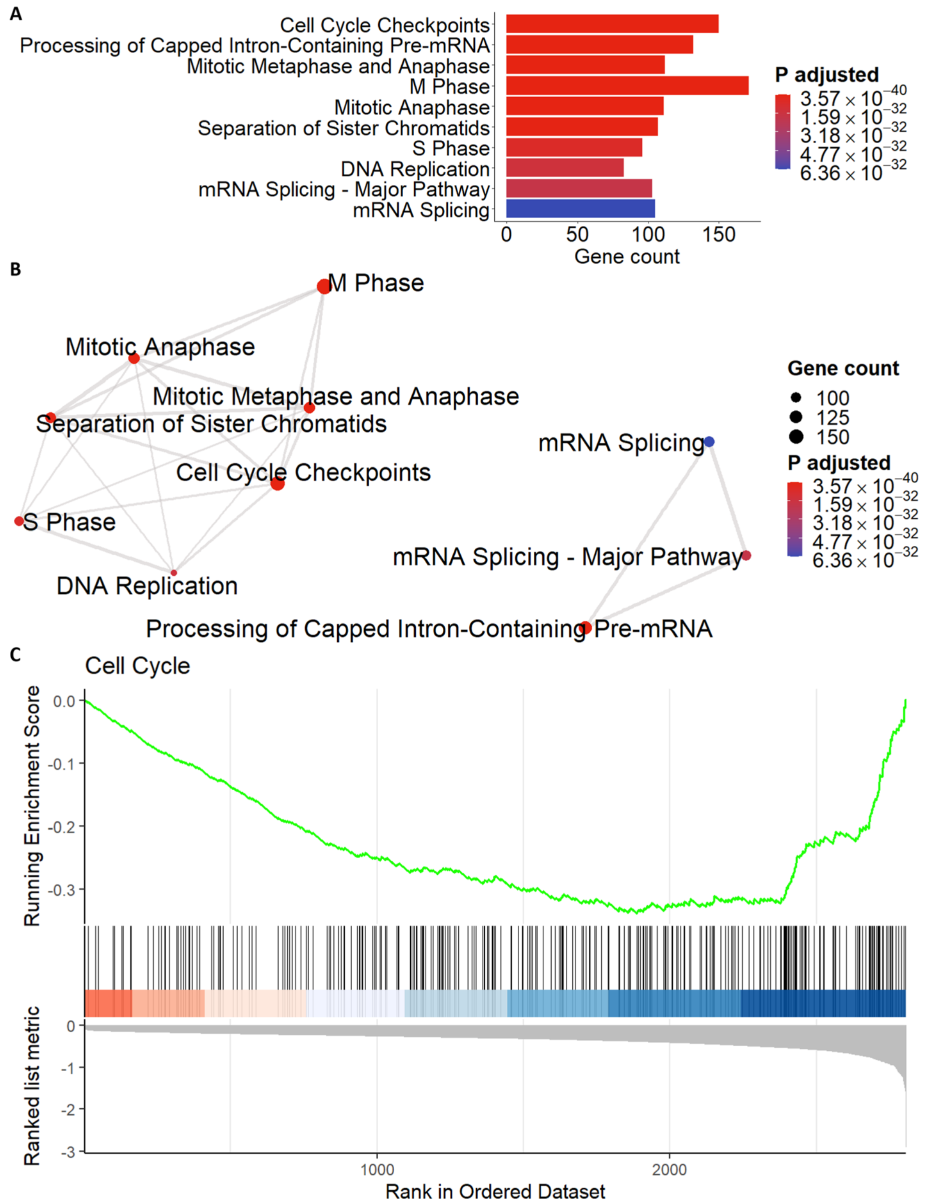


Fig. 6 Identification of enriched pathways among differentially expressed genes. **a** Bar graph showing the results of over-representation analysis using the `Pathway Enrichment Plot` tab. The maximum number of pathways/categories to show was set to 10 and the enrichment p_{adj} value cutoff was set to < 1×10^{-30}. The gene count (*x*-axis) indicates the number of genes enriched in each pathway and colors indicate level of significance (p_{adj}). The pathways are plotted on the *y*-axis in order of increasing significance. **b** While the `Pathway Enrichment Plot` tab shows a bar graph or dot plot for only a subset of enriched pathways, the `Pathway Enrichment Map` tab shows all of the enriched pathways in an interconnected network map. The size of each node indicates the gene count (number of genes enriched in each category) and the color represents the p_{adj} value (the cutoff was set to < 1×10^{-30}). **c** The `GSEA Plot` tab generates a plot of the running enrichment score for a specified pathway/category. The plot for the category "cell cycle" is illustrated here. Currently only Reactome pathways/categories are supported for each of these figures. The input data for **a**-**c** is the filtered or unfiltered data from the `Gene Table` tab (we set the LFC to < 0 and the p_{adj} cutoff to < 0.05)

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: Log_2 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

[1]London Regional Cancer Program, London Health Sciences Centre, London, ON N6A 5W9, Canada. [2]Department of Biochemistry, University of Western Ontario, London, ON N6A 5C1, Canada. [3]Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON N6A 5C1, Canada. [4]Children'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-1α-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.