DEVEA: an interactive shiny application for Differential Expression analysis, data Visualization and Enrichment Analysis of transcriptomics data

Miriam Riquelme-Perez, Fernando Perez-Sanz, Jean-François Deleuze, Carole Escartin, Eric Bonnet, Solène Brohard

SOFTWARE TOOL ARTICLE

REVISED DEVEA: an interactive shiny application for Differential Expression analysis, data Visualization and Enrichment Analysis of transcriptomics data [version 2; peer review: 2 approved]

Miriam Riquelme-Perez[1,2*], Fernando Perez-Sanz[3*], Jean-François Deleuze[2], Carole Escartin[1], Eric Bonnet[2*], Solène Brohard[2*]

[1]Université Paris-Saclay, CEA, CNRS, MIRCen, Laboratoire des Maladies Neurodégénératives, Fontenay-aux-Roses, 92265, France
[2]Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, 91000, Evry, France
[3]Biomedical Informatics & Bioinformatics Service, Institute for Biomedical Research of Murcia (IMIB), Murcia, 30120, Spain

* Equal contributors

v2 **First published:** 28 Jun 2022, **11**:711
https://doi.org/10.12688/f1000research.122949.1
Latest published: 24 Mar 2023, **11**:711
https://doi.org/10.12688/f1000research.122949.2

Abstract

We are at a time of considerable growth in transcriptomics studies and subsequent *in silico* analysis. RNA sequencing (RNA-Seq) is the most widely used approach to analyse the transcriptome and is integrated in many studies.
The processing of transcriptomic data typically requires a noteworthy number of steps, statistical knowledge, and coding skills, which are not accessible to all scientists. Despite the development of a plethora of software applications over the past few years to address this concern, there is still room for improvement.
Here we present DEVEA, an R shiny application tool developed to perform differential expression analysis, data visualization and enrichment pathway analysis mainly from transcriptomics data, but also from simpler gene lists with or without statistical values.
The intuitive and easy-to-manipulate interface facilitates gene expression exploration through numerous interactive figures and tables, and statistical comparisons of expression profile levels between groups. Further meta-analysis such as enrichment analysis is also possible, without the need for prior bioinformatics expertise.
DEVEA performs a comprehensive analysis from multiple and flexible data sources representing distinct analytical steps. Consequently, it produces dynamic graphs and tables, to explore the expression levels and statistical results from differential expression analysis. Moreover, it generates a comprehensive pathway analysis to extend biological

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 24 Mar 2023		✓ view
version 1 28 Jun 2022	✓ view	? view

1. **Hélène Hirbec**, Université de Montpellier, Montpellier, France

2. **Wenbin Guo**, James Hutton Institute, Dundee, UK

Any reports and responses or comments on the article can be found at the end of the article.

insights. Finally, a complete and customizable HTML report can be extracted to enable the scientists to explore results beyond the application. DEVEA is freely accessible at https://shiny.imib.es/devea/ and the source code is available on our GitHub repository https://github.com/MiriamRiquelmeP/DEVEA.

Keywords
Bioinformatics, transcriptomics, RNA sequencing, differential expression analysis, enrichment analysis, visualization, R, Shiny, interactive reports.

This article is included in the Bioinformatics gateway.

Corresponding authors: Miriam Riquelme-Perez (miriam.riquelmep@gmail.com), Fernando Perez-Sanz (fernando.perez8@um.es), Eric Bonnet (eric.bonnet@cnrgh.fr), Solène Brohard (solene.brohard@cnrgh.fr)

Author roles: Riquelme-Perez M: Conceptualization, Data Curation, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Perez-Sanz F**: Conceptualization, Data Curation, Resources, Software, Validation, Writing – Review & Editing; **Deleuze JF**: Funding Acquisition, Resources; **Escartin C**: Supervision, Validation, Writing – Review & Editing; **Bonnet E**: Supervision, Validation, Writing – Review & Editing; **Brohard S**: Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: MRP holds a PhD fellowship from the CEA (Amont-Aval).

How to cite this article: Riquelme-Perez M, Perez-Sanz F, Deleuze JF *et al.* **DEVEA: an interactive shiny application for Differential Expression analysis, data Visualization and Enrichment Analysis of transcriptomics data [version 2; peer review: 2 approved]** F1000Research 2023, **11**:711 https://doi.org/10.12688/f1000research.122949.2

First published: 28 Jun 2022, **11**:711 https://doi.org/10.12688/f1000research.122949.1

REVISED **Amendments from Version 1**

This second version of DEVEA has included a new plant reference species (*Arabidopsis thaliana*) as a major change.

Some recommendations and additional explanations in the main text make it easier to read and clarify some parts of the article including good practices (for more in-depth discussion, please check the reviewers' responses).

New supplementary material has been included in the GitHub repository where the application's source code is located, which is recommended to be explored if required.

Lastly, some application bugs have been fixed in order to improve the experience of using DEVEA. Nevertheless, do not hesitate to contact the authors if you encounter any error in the use of the tool.

Any further responses from the reviewers can be found at the end of the article

Introduction

RNA sequencing (RNA-seq) has become a routine and popular technique for genome-wide and transcriptomics expression analysis.[1] As a result, techniques analyzing RNA are extensively incorporated in basic science research and are even increasingly used as molecular diagnostics for human health. These may include diagnosis, prognosis and therapeutic selection.[2]

However, in order to leverage the full power of this technique, several stages and tools are necessary to translate expression profiles into valuable outcomes. The R statistical environment[3] provides many well-known packages to perform key steps of a complete RNA-seq analysis pipeline. Possible examples include differential expression analysis (DEA) functions, leading to lists of differentially expressed genes (DEGs), and annotation enrichment analysis (EA, sometimes called pathway analysis) libraries, which will identify biological pathways or cellular functions significantly enriched from the list of DEGs.

Nevertheless, most of these powerful packages are command-line based or demand coding knowledge and are therefore out of reach for scientists with limited computational training. Besides, analyses can be started at different points in the workflow, from raw or partially analyzed data from different tools, to individual lists of favorite final features. Thus, tools providing several ways to start an analysis are more flexible than others using a single data type. Providing flexible user-friendly tools for the analysis and visualization of gene expression data can help researchers to move from high-throughput genomics to basic scientific research. To bridge this gap, an increasing number of software tools are being released, based on intuitive, point-and-click, graphical interfaces. Frameworks such as R Shiny,[4] an R package, facilitate the creation and release of interactive web tools. Certain RNA-seq analysis applications from the literature may include iDEP,[5] GENAVi[6] and ideal,[7] among others.

However, there are still ways to improve the functionalities of these tools. For instance, supporting input data types of different levels of complexity can extend the level of performance of the tool. It is also essential to include widely used types of analyses such as differential expression and pathway analysis, with enough options for calculations and graphical representation to generate valuable results. At last, user-friendliness is a particularly important point since these tools aim at helping the non-specialist. Therefore providing a robust and easily accessible web interface is an essential asset. With these considerations in mind, we have developed DEVEA, a new interactive R Shiny application for DEA, data exploration, data visualization and functional EA. DEVEA provides an easy-to-use interface to load data in various formats and complexities according to the stage of the analysis, including raw RNA-seq count data, pre-analyzed data, simple lists of genes or proteins obtained from different sources, with or without statistical values associated. From these different types of input data, it generates a wide set of dynamic plots and tables allowing quick navigation through gene expression profile or enrichment analysis results. The outputs can be downloaded easily and the user can create custom and operable reports in HTML format. DEVEA is implemented as a publicly available web server and can be optionally downloaded to be used locally. DEVEA aims to conduct a proper analysis by reaching out to both life scientists (gathering the biological expertise) and bioinformaticians (offering the technical expertise), and to foster communication between the two sides to promote easier and more extensive analysis of data.

Methods

Operation

DEVEA was built as a Shiny application[4] in R[3] (V.4.1.1). Shiny is a package that facilitates the development of web applications from R. It is particularly indicated for building interactive and user-friendly software wrappers.

The tool is hosted on a remote, freely accessible web server (http://shiny.imib.es/devea). Apart from DEVEA's public web server, the application can be used on a local computer (see the supplementary material for a detailed procedure here https://github.com/MiriamRiquelmeP/DEVEA/blob/main/Supplemental-Information.md). Its source code is available on GitHub (https://github.com/MiriamRiquelmeP/DEVEA), under the terms of the Apache license 2.0. DEVEA has been tested in Linux and Windows 10 operating systems locally, and has also been launched remotely with different browsers (Google Chrome, Mozilla Firefox and Internet Explorer). However, for the best user experience in terms of rendering and visualization, we recommend to use Mozilla Firefox. Other browsers may present display issues when deploying some of the elements of the application and generate errors. Locally running the application shares all the same characteristics as the Shiny web application. A comprehensive guide on how to use the application from the different input modes can also be explored through the accessible tutorial from both DEVEA modules (DESeq DEVEA and Simple DEVEA) in the 'Tutorial' section from the top controls and independently on https://shiny.imib.es/DESeqDevea/tutorial.html or https://shiny.imib.es/simpleDevea/tutorial.html.

DEVEA relies on several existing R packages to carry out all the functionalities proposed (see supplementary material for the complete list). For instance, in order to handle the calculation of DEGs, the analysis is largely based on DESeq2 package.[8] The annotation is managed by the AnnotationDbi package,[9] collecting all dedicated annotation databases for the different species (currently *Homo sapiens, Mus musculus, Rattus norvegicus* and *Arabidopsis thaliana*) for robust name conversion. For the enrichment calculations and visuals, the R packages topGO,[10] fgsea (Fast Gene Set Enrichment Analysis)[11] and clusterProfiler[12] together with other basic dependencies, such as ggplot2[13] and plotly[14] have been used. All those packages are widely used by the community and well maintained. We have selected DESeq2 for differential expression analysis as it is one of the most popular and well-tested tools for this task. Furthermore, it was shown[15] that DESeq2 has a good overall performance regardless of presence of outliers and proportion of DE genes compared with other methods, such as Limma-Voom,[16] BaySeq[17] and edgeR[18] among others.

The full DEVEA global workflow is shown in Figure 1. The main analysis path can be launched at different steps depending on the four input modes (check *Data requirement section* for details), represented at the top of the figure. The dashed arrows indicate where every input is incorporated in the pipeline, until the end of the workflow. Objects that are more complex will generate more results. For each step of the analysis, intermediate results are available as tables and graphical representations in their dedicated spaces, detailed below in the circular notes. The vast majority of tables and plots are interactive, allowing the user to visualize data in real-time as well as to interact efficiently and can be individually downloaded. In the end, a global report can be generated and annotated. Each of these steps of the complete analysis will be described in their corresponding sections in the next paragraphs.

Data requirement

The tool has four main data input modes (Figure 2):

- **CM + SI mode:** refers to a **counting matrix (CM)**, containing the raw number of counts per gene as round digits, where columns correspond to samples and rows to features. Only un-normalized raw counts are accepted as input data, as obtained from the counting process. Data from other sources containing decimals (e.g. RSEM) has to be rounded before being uploaded in DEVEA. The CM should be associated with another file gathering the **sample information (SI)**, as a data frame containing metadata about each sample, with the first column the identifier used by default as a label in visualizations. This can be modified afterward during the analysis. It should include any other relevant experimental factor (e.g. treatment/control, sex, cell type, tissue, etc). The design of the comparison will be determined by one of these factors. The column names in the CM and the first row names in the SI must be identical, and the gene IDs in this file can be included in Symbol or ENSEMBL format. Both files can be in .CSV, .TXT or .XLSX format.

- **DO mode:** based on a **DeseqDataSet object (DO)** generated by the *DESeq()* function from DESeq2 package. It is an object used to store the input values, intermediate calculations and results from a DEA. The user must have created it with the *CountData* field as the data matrix of counts, the *ColData* field with the sample information and a design formula specifying the experimental level to test for DEA. The first column in the CountData and the first row in the ColData are equal. The gene names can be included as gene Symbols or ENSEMBL gene IDs. The object must be compressed and extracted from R in .RDS format. If the differential expression object has been generated with a different tool or package, you may use the DEFormats[19] R function for a possible conversion.

- **GL + SV mode:** a **gene list (GL)** with associated **statistical values (SV)** per gene. The first column should contain gene names (in Symbol or ENSEMBL format), the second column the fold-change and the third column the statistical adjusted p-value, in this precise order. Column names should be provided without special

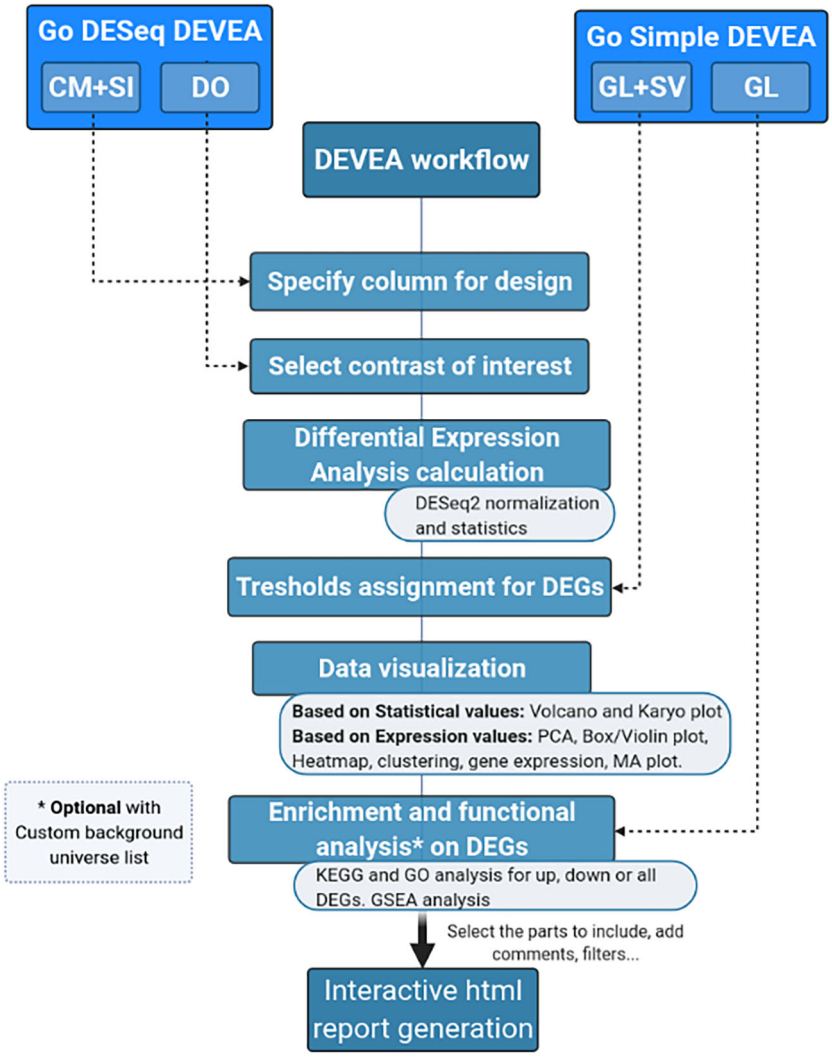

Figure 1. DEVEA global workflow.

characters. The numerical values will be used without further modifications by DEVEA to set the threshold of expression change and the significance. This table can be uploaded in .CSV, .TXT or .XLSX file format.

- **GL mode:** a **Gene list (GL)** consisting on a single column file in .CSV, .TXT or .XLSX format containing the gene names (in Symbol or ENSEMBL) and including a column name without special characters (i.e. GeneName, Genes, ID, etc). The gene list can be copied and pasted directly into the dedicated field in DEVEA, without the column name.

While the main data type for DEVEA's usage is RNA-seq data, it is worth noting that the simple gene list (GL) can be built from any other type of "omics" datasets, as long as the identifiers are recognized by DEVEA. Another example could be the use of GL + SV mode with treated data from microarray analysis, where values such as FC of expression between groups and adjusted p-value are available from the signal intensity of the probes. It is highly recommended to work with log2FC and adjusted p-values. The more elaborated input data types, such as the counting matrix (CM + SI), can be built in some cases from different data where they can safely be processed by DESeq2 functions. An example may be mass spectrometry data, the method of choice for quantitative. With label-free proteomics, it is possible to quantify proteins by using their spectral counts as an approximation of protein abundance, and then use statistical models such as DESeq2 even if they are designed specifically for count data. A study comparing different statistical methods for differential expression detection in label-free mass spectrometry proteomics shows that DESeq2 performed well both in terms of detection of true positives as well as controlling for the number of false negatives.[20] Therefore, it is perfectly possible to

CM+SI

ID	Ctrl_1	Ctrl_2	Ctrl_3	Ctrl_4	Trt_1	Trt_2	Trt_3	Trt_4
FeatureA	57	41	64	55	38	45	31	39
FeatureB	71	40	100	81	41	77	58	40
FeatureC	256	177	220	189	107	213	172	126
FeatureD	0	1	1	0	0	0	0	0
FeatureE	146	81	138	125	52	91	80	50
FeatureF	10	9	2	5	2	9	8	4
FeatureG	3	2	6	5	2	2	1	0
FeatureH	0	0	2	1	3	0	2	1
FeatureI	56	37	107	118	65	49	52	37

+

Sample	Treatment	Sex
Ctrl_1	Control	Male
Ctrl_2	Control	Male
Ctrl_3	Control	Female
Ctrl_4	Control	Female
Trt_1	Treatment	Male
Trt_2	Treatment	Male
Trt_3	Treatment	Female
Trt_4	Treatment	Female

DO

```
library('DESeq2')

dds <- DESeqDataSetFromMatrix(
  countData = CM,
  colData = SI,
  design = ~Sex )

DO <- DESeq()

saveRDS(DO, file = "DO.rds")
```

GL+SV

ID	FC	p-value
FeatureA	-0.189	0.000488
FeatureB	1.427	1.998E-05
FeatureC	-2.157	3.500E-07
FeatureD	-3.558	0.0058756
FeatureE	0.844	0.0004858
FeatureD	-3.558	0.0058756
FeatureF	1.665	8.045E-06
FeatureG	3.564	2.389E-05
FeatureH	-6.574	0.0007451
FeatureI	0.012	0.0027583

GL

ID
FeatureA
FeatureB
FeatureC
FeatureD
FeatureE
FeatureD
FeatureF
FeatureG
FeatureH
FeatureI

Figure 2. Possible data input formats for DEVEA. 'Feature' represents a Gene Symbol or ENSEMBL gene ID.

use this type of data in DEVEA, as long as the values represent unique measurements as integer numbers, and the protein IDs are replaced by their coding gene name.

Implementation

Getting started

To start working with DEVEA, the adequate module to perform the analysis has to be chosen from DEVEA's main lobby interface. The decision depends on the input data format. The user has to choose 'Go DESeq DEVEA' if their input is a CM + SI or a DO, and the 'Go Simple DEVEA' mode in the case of a GL + SV or a simple GL input files. See Figure 3 for a visual screenshot of the lobby.

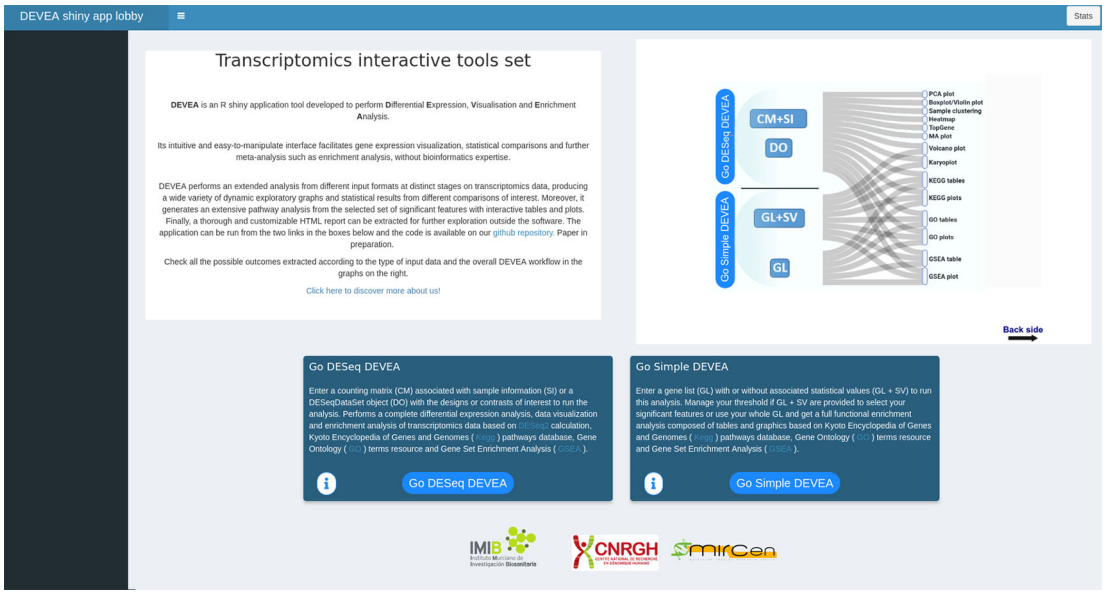

Figure 3. Generic screenshot of the lobby app showing the two possible pathways to start the analysis.

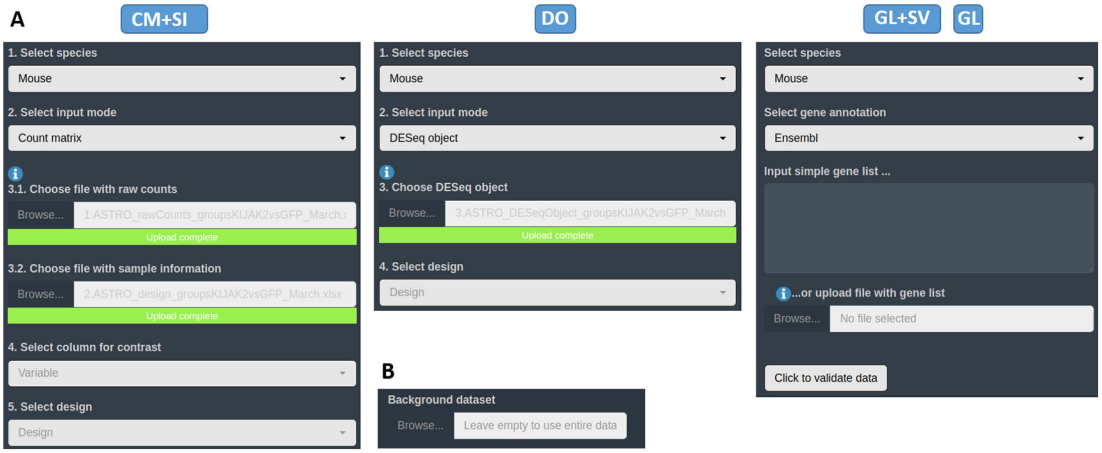

Figure 4. DEVEA's import section. A: Upload spaces for different types of input data. B: Section to specify a custom background universe for enrichment analysis.

Data upload and statistical design specification

Within the appropriate interface according to the data type, the first tab available corresponds to the *Input data section.* The user has to upload their own data in one of the different accepted formats and types (see them on *Data requirement section*) in their dedicated spaces (see Figure 4A). For all input data, a field to specify the custom dataset to use as a background universe is available as well (see Figure 4B). If necessary for the user, some *demo data* representing the four different input types, can be found on DEVEA's GitHub https://github.com/MiriamRiquelmeP/DEVEA/tree/main/data. The nature of the *demo data* and how it was generated are described in the *Use case section.*

When a CM + SI or a DO is used as input data, it is important to indicate the statistical design or contrast for the expected comparison. The design formula expresses the variables that will be used to calculate the differential expression in following steps. For the CM + SI input format, the levels of interest that will compose the final design must be included in one of the columns of SI file. By entering the column name in the 'Select column for contrast' field, the program extracts the conditions and calculates the combination based on the distinct levels (i.e. *Treatment_Control_vs_Treatment* if column Treatment is selected or *Sex_Male_vs_Female* will be displayed if Sex is selected from SI in Figure 2 - CM + SI). In cases where more than two levels are available, the application will propose all one-vs-one combinations. Then, the relevant combination for the analysis has to be selected by the user in the 'Select design' part. In a DO, the user can use the function *relevel()* from DESeq2 in R to specify the basal level.

With a DO data type, the column for the design must have been already incorporated when generating the DESeq2 object in R. Only simple designs will be generated and/or can be selected from 'Select design' field (i.e. *Sex_Male_vs_Female* if *design = Sex* is specified in the formula as in Figure 2 - DO). Several conditions to be treated can be included in the design of DESeq2 within the same DO. DEVEA will offer the possibility to consider any of them for each analysis, and the user can select which contrast to explore from the same DO.

It is possible to model batch effects in DEVEA with the DO object. In that case, the user can include the batch effect directly in the design formula (i.e. *design = ~ Batch + Condition*). Once the final contrast is specified as *Condition_level1_vs_level2*, the batch effect will be accounted for in the statistical model. In that case, the batch effect is treated in DEVEA as a covariate in the regression model.

The user can also remove the batch effect before uploading the DO object into DEVEA, for instance by using the *removeBatchEffect()* function of the Limma-Voom R package,[16] or other packages such as ComBat-seq.[21]

Differential expression analysis (DEA) and data view

The first key performance of the application consists in extracting the descriptive information based on the feature expression and the statistical contrast for DEA. In CM + SI data type a new DeseqDataSet object is calculated from the files and information provided by the user. In DO or GL + SI input modes, the application retrieves the important values already included in these objects. All transformations, normalization and measurements applied to the data at this step are

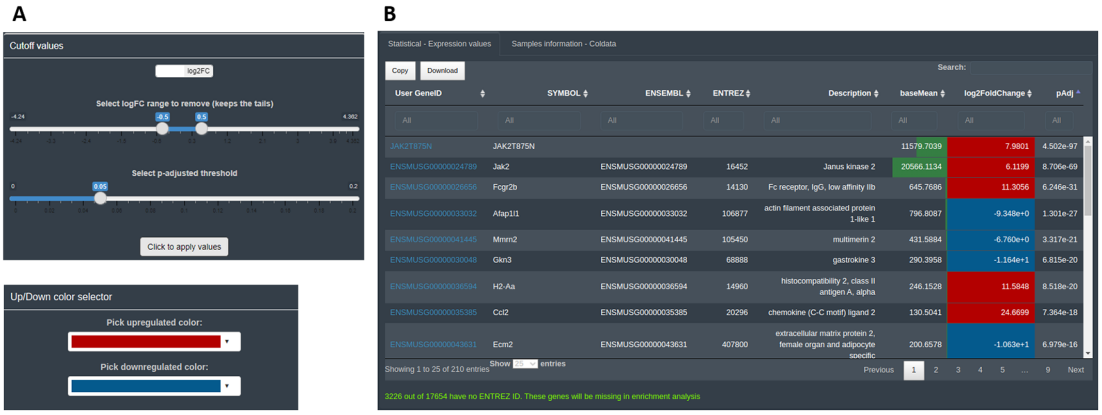

Figure 5. Basic data visualization for DEGs in DEVEA. A: Controls used to set the statistical thresholds and graphical features parameters. B: Interactive result table.

performed with functions included in DESeq2 R package. It should be noted that DEA calculation is not possible with the simple GL input mode, due to the lack of expression values and statistical details. The following comments will not apply to this object.

The calculations and the statistical results are accessible in the 'Preview dataset' section tab. At this step of the analysis, the user can explore the number of features considered as differentially expressed and their direction, and establish the descriptive statistics thresholds to consider them DEGs. By default, DEVEA uses prefixed log2 fold change $|lfc| > 0.5$ and adjusted p-value < 0.05 thresholds, that can be adapted by the user at any moment and thus modulate the list of DEGs. Moreover, the information uploaded and the descriptive statistics will be used to establish and control some interactive parts of the plots. For example, the color of defined up-regulated or down-regulated genes can be chosen (Figure 5A). For CM + SI and DO, a complete table of results, named 'Statistical - Expression values', is displayed showing useful information such as base means across samples, log2 fold changes, standard errors, raw and adjusted p-values for the specific design selected. A second table is also shown with the DEA details called 'Samples information - Coldata' (Figure 5B). For the GL + SV mode, raw data are displayed in a table. The user can monitor gene name conversion, explore values interactively and sort, filter and download them at any time.

It is important to stress that, as different elements can be extracted from the distinct input objects depending on their complexity, the number of graphs available for each of them vary (Figures 1 and 6). Using the raw expression values that are available only in CM + SI and DO input modes, the user can explore data in a Principal Component Analysis plot (PCA) with the top 500 variant features, to show clusters of samples based on their similarity selecting the principal components of interest; a box or violin plot for gene expression distribution across the dataset; a heatmap representation

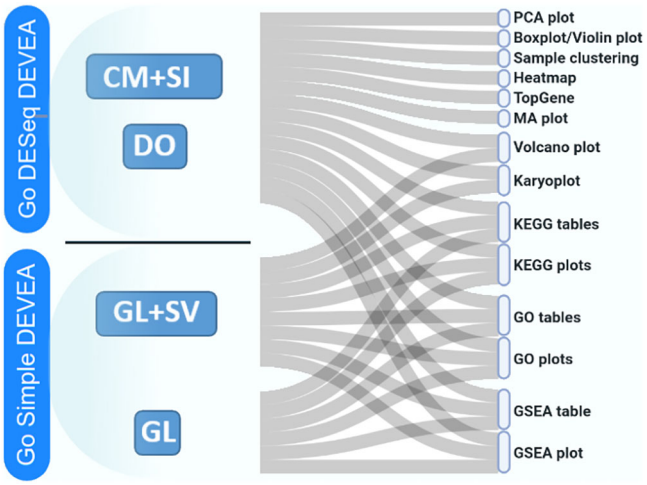

Figure 6. Possible types of graphical representations in DEVEA depending on data input type.

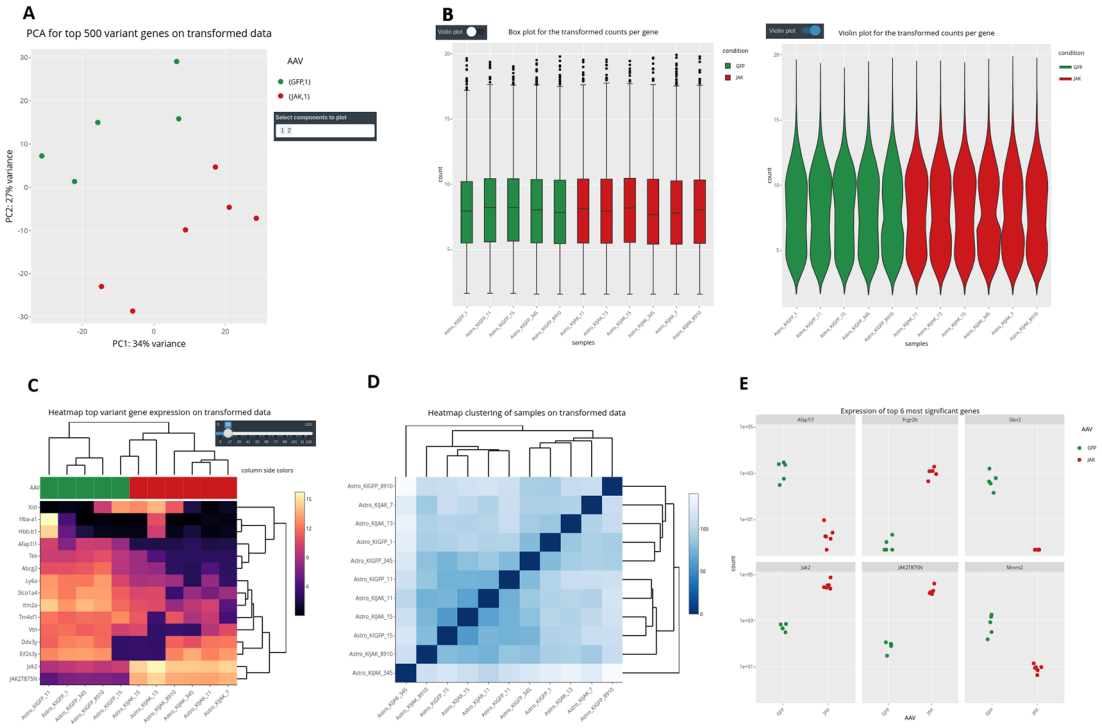

Figure 7. Advanced data visualization in DEVEA. A: PCA plot. B: Box and violin plots. C: Gene expression heatmap. D: Sample hierarchical clustering. E: Dot plots.

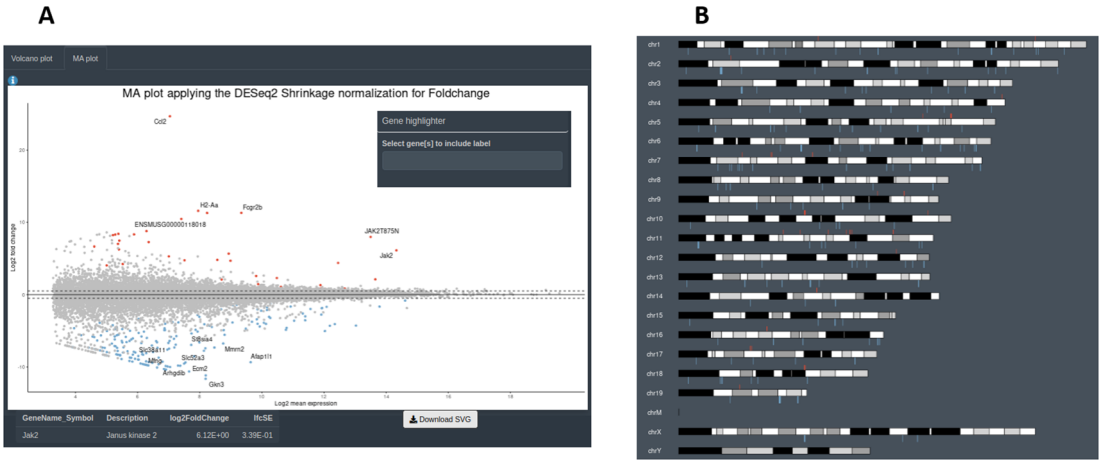

Figure 8. Additional data visualizations available from DEVEA. A: Section to explore volcano and MA plots. B: Karyotype plot.

of the top variant genes, regulated by the user; and a dot plot with the expression of the top 6 variant genes or a selection of individual genes of interest (Figure 7). A second group of graphs represents genes related only to their statistical values. This can be displayed from CM + SI, DO and GL + SV modes. They consist in a volcano plot, that shows statistical significance (adjusted p-value) versus the magnitude of change (FC) regarding the contrast levels; and a karyotype plot showing the DEG position on the genome and the direction of change (color coded by up- or down-regulated) (Figure 8B). Finally, it is also possible to combine gene expression with statistical values, from CM + SI and DO input modes, to generate a MA plot (an application of a Bland–Altman plot for visual representation of genomics data[22]) (Figure 8A) displaying feature labels and statistical values by clicking on each gene dot.

Enrichment analysis (EA) and visualization

The last stage of the analysis with DEVEA is EA. This is a method to identify classes of defined categories that are over-represented in the list of DEGs. These categories may be associated with disease phenotypes, biological pathways, or cellular functions. DEVEA uses the differentially expressed and significant features to retrieve the over-represented terms from several well-known databases. This major block of the DEVEA analysis can be carried out from all data input types. It consists of an extensive EA after the selection of appropriate statistical values for defining DEGs from CM + SI, DO and GL + SV, or using all components included in the simple GL input. It collects significant terms from KEGG (Kyoto Encyclopedia of Genes and Genomes[23]) and GO (Gene Ontology[24]) Biological process, Molecular function and Cellular component databases. Furthermore, a GSEA (Gene Set Enrichment Analysis[25]) and leading edge exploration analysis can be performed from different databases for the whole set of features. In the case of CM + SI, DO and GL + SV input modes, KEGG and GO analyses are performed for all DEGs together, and for the subset of up- and down-regulated genes in separated tabs. GSEA is always performed on the complete set of genes and uses the statistical values associated with the features. With the simple GL data input, enrichment can only be performed for the whole set of genes for KEGG and GO analysis and no GSEA will be possible, due to the lack of statistical information.

The main results of EA are shown as interactive tables containing detailed information on the enrichment from each database. In KEGG and GO categories, the tables display columns for the name of the significant pathways or terms, their adjusted p-values and additional descriptive information such as total number of genes associated or the DEGs participating in the pathway. The user can also display the gene names that match in the pathway from the "+" symbol. Below each table, additional plots can be created by selecting rows of interest (showed in green on the Figure 9A). The plots are interactive and reactive. They can be changed at any time by selecting new lines in the table. The user can visualize results as word cloud, circle plot, bars plot, chord plot, dot plot, heatmap or net plot representing different elements from the tables. For GSEA, the results are displayed as a table containing the significant enriched pathways from the selected databases. In the table, important GSEA calculation parameters are available, as leading edge analysis. This allows determining which subsets of genes contributed the most to the enrichment signal of a given pathway. Below, a typical GSEA plot is shown from the lines selected in the table (see Figure 9B). Leading edge results have also been implemented in the plot when one unique pathway is displayed, as a red line indicating the extent of what we consider the leading edge genes.

As a special feature offered by DEVEA, a custom gene list can be uploaded to be used as the background gene universe for EA (Figure 4B). For example, the user can use a background list containing only expressed genes in this experiment. It will control for experimental context and enable representative functions, rather than only the functions explaining the nature of the samples, to be identified. This is especially important to consider in microarray studies when a limited number of probes is used, or in studies with specific cells or tissues showing a restricted set of detectable genes.

For this EA, internal ENTREZ ID gene code is used to associate gene names in Symbol or ENSEMBL with the enrichment annotation in KEGG, GO and GSEA libraries. An exhaustive conversion across different functions is

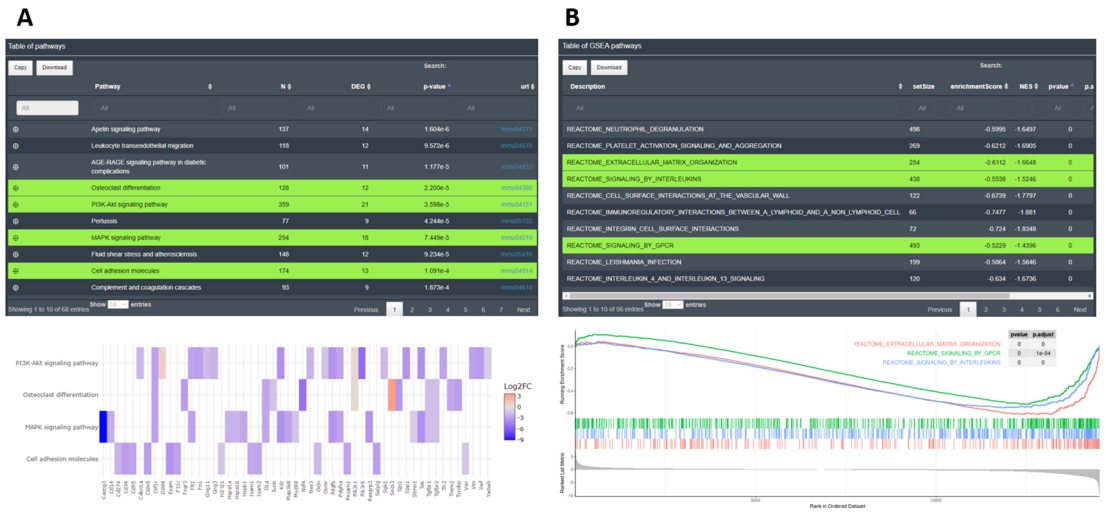

Figure 9. Examples of enrichment analysis with DEVEA. A: KEGG EA table and heatmap. B: GSEA table and plot.

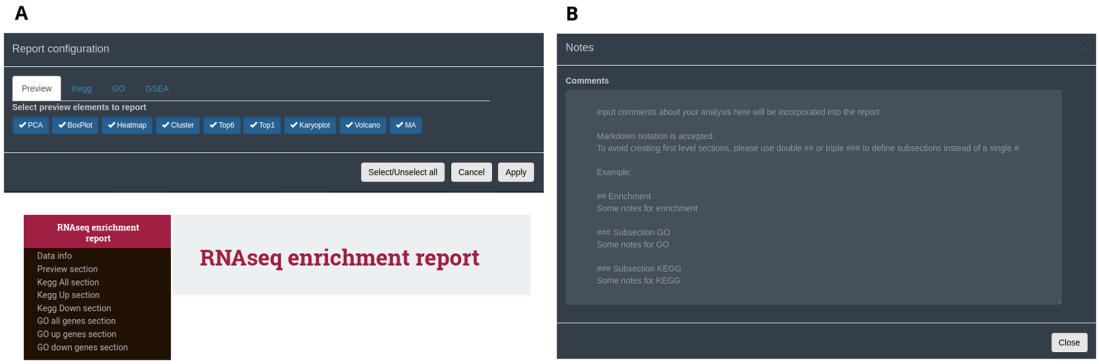

Figure 10. Interactive final DEVEA report functionalities and features. A: Configuring box and final HTML report. B: Notes transferable to the report.

conducted to retrieve all possible terms, since they are not homogenously registered in all conversion databases. Furthermore, not all genes are curated or annotated and therefore some will be left out of the EA (the portion of genes is indicated in the application below the 'preview table' as shown in Figure 5B). This is a limitation from the automated databases conversions without manual curation. For this reason, the number of species is limited and not all genes will be used for the EA, but the results obtained are more robust.

Global report

An interactive HTML report can be generated from all data types following analysis and exploration with DEVEA. It is available in the 'HTML report' button at the top fixed part of the application. To create it, the user can select the individual set of figures, tables and results to be kept in this single HTML document (Figure 10A). Plots will retain the last aesthetic indicated in the graphical parameters (e.g. colors, shapes, labels, terms). Only full tables will be kept, without taking into account potential filters applied during the analysis, allowing full exploration, sorting and re-filtering of the whole dataset outside of DEVEA. It should be noted that the majority of the results can also be copied or downloaded in high-quality format at any step of the analysis within DEVEA. Importantly, comments can be included at any step of the analysis in a dedicated section at the top right position of the application, and will be automatically saved (Figure 10B). They can be displayed in the final report to ensure that the observations made throughout the analysis, with special interpretations or results are maintained.

Please note that if any of the graphs or tables fails to be generated in the application, when there are not enough results, genes or pathways to display them, the report cannot be generated. Unselect the problematic plots or tables to be able to render the rest of the report without errors.

Use case

To demonstrate the usefulness of DEVEA, an RNA-seq dataset generated by our team and published recently[26] was investigated using the application. The study aimed to characterize the role of JAK2-STAT3 signaling in astrocytes in the context of Huntington's disease (HD). HD is a rare genetic neurodegenerative disease leading to severe motor, cognitive and psychiatric symptoms, with no curative treatment available.[27] Astrocytes, a heterogeneous group of star-shaped glial cells, perform key functions in the brain. They provide nutrients to neurons, regulate synaptic transmission, and contribute to brain repair following injury.[28] Astrocytes become reactive in the brain of HD patients and their impact on HD progression is still unclear.[29] The study used a genetic mouse model of HD. Treated mice were injected with an adeno-associated viral (AAV) vector targeting astrocytes and encoding a constitutive form of the JAK2 kinase (JAK2T875N) to activate the JAK2-STAT3 pathway. Control mice were injected with a similar AAV expressing GFP. Astrocytes were isolated and sequenced by RNA-seq on a HiSeq 2500 Illumina platform (2×100 bp). Quality control of sequencing data was performed with FastQC[30] (v0.11.9). Reads were mapped on the GRCm38 (mm10) mouse genome assembly with Hisat2[31] (v2.2.1), and a counting matrix was generated. Quantification of reads associated with genes was achieved with featureCounts[32] (v2.0.0), and differential gene expression analysis was performed with DESeq2 (v1.28.1) Bioconductor (v3.13) package on R (v4.0.2). Only genes with a raw number of counts ≥ 10, in at least 3 samples were analyzed. Data were adapted and integrated as different input objects in DEVEA to test all functionalities. For instance, Figure 7A shows that control (GFP, N = 5, in green) and treated (JAK, N = 6, in red) samples are clearly separated on a PCA plot, with a better separation achieved on PC2 (representing 27% of the total variance). Figure 7C- D also show two types of clustering profiles, which group samples from the same group together and display genes with higher variability. Figure 11A shows that the levels of *Jak2* are higher in treated (JAK) versus control (GFP) groups, as

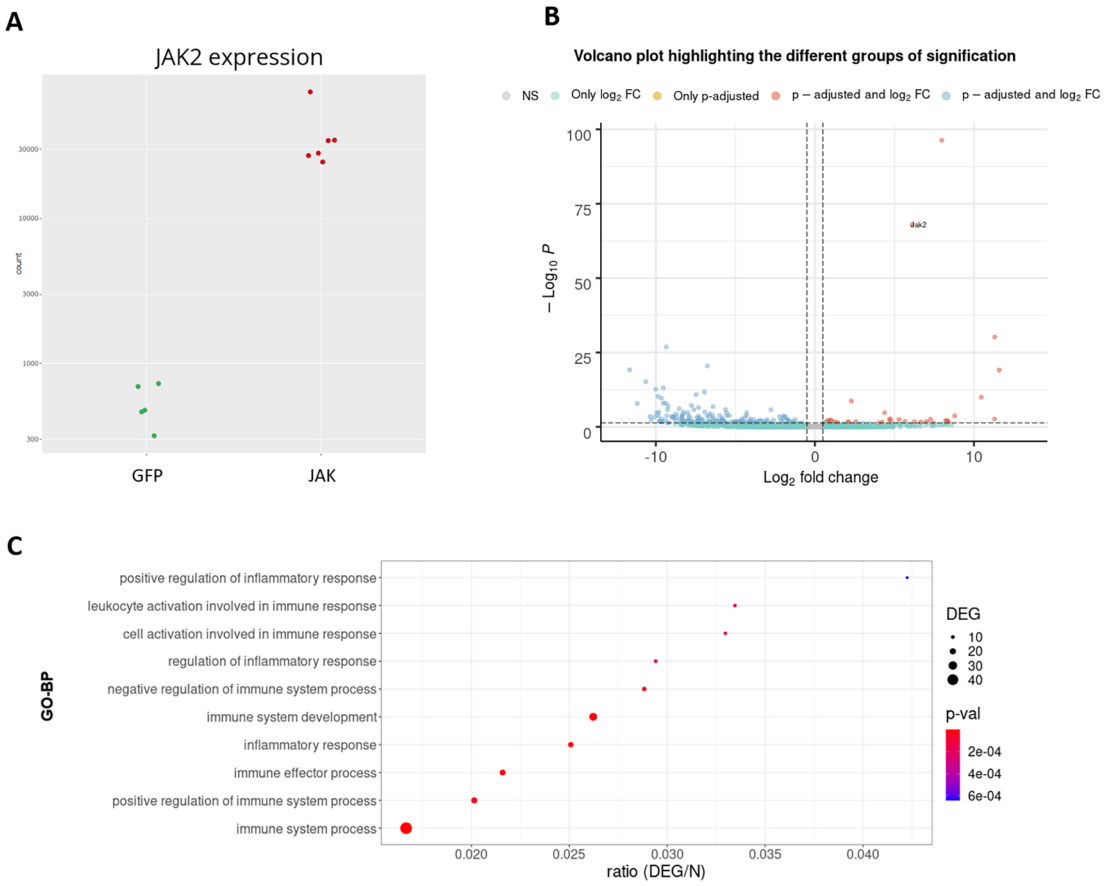

Figure 11. RNA-seq analysis of the JAK2-STAT3 pathway in astrocytes of a Huntington's disease model mouse. A: Expression levels of the *Jak2* gene between groups. B: Volcano plot. C: GO-BP enriched categories.

expected by AAV-mediated gene transfer (log2(FC) = 6, associated adjusted p-value 8.7E-69). In addition, a volcano plot demonstrates that JAK2 causes down-regulation of many genes, shown in blue in the upper left quartile of the graph on Figure 11B. Finally, EA in the treated (JAK) versus control (GFP) list of DEGs shows that many GO-BP terms are related to Immunity/Inflammation, a process linked to the reactive changes in astrocytes induced by JAK2 (Figure 11C).

All data corresponding to this research project are available under the four different DEVEA input types. They are available on the GitHub web site (see the *Software Availability section*). They consist in (1) a set of CM + SI, where features are genes in the ENSEMBL format and expression represents raw counts after alignment and filtering to remove non-expressed genes from the CM. The final number of features is 18,260 (+ 2 custom genes representing the two transgenes *Jak2T875N* and *Gfp*). The SI file contains all relevant information on sample characteristics; (2) a complete DO built from the same CM and SI files with a design based on the comparison of the two different AAVs (design = AAV); (3) a GL + SV file manually created according to the *DESeq2* results. The file is generated with 3 columns that contains the gene names, FC and p-values for the top 500 most significant genes and (4) a unique GL containing only the 268 DEGs from the comparison of interest, reported from the DESeq2 analysis with adjusted p-value < 0.1 statistical threshold and no threshold for the FC (for further details on the input data, see the *Data requirement section*).

With this user case, no errors were detected throughout the analysis. DEVEA, thanks to a large range of graphical and statistical analyses, highlighted significant differences between reactive astrocytes in the JAK group and control astrocytes in the GFP group. Due to extensive EA, DEGs were associated with important biological functions such as immunity/inflammation pathways as well as cytokine signaling and proteostasis. These results are consistent with the conclusion described in the corresponding publication.[26]

We have also generated demo files corresponding to an RNA-seq study of *Arabidopsis thaliana* from a recent publication. The data represents the differential expression analysis of 3 mutant fibrillins (FBN6) samples vs 3 controls. See the dedicated paper for further details.[33]

Table 1. DEVEA functionalities compared to similar software tools in their online versions.

	DEVEA	iDEP[5]	ideal[7]	GENAVi[6]	RNfuzzyApp[37]	DEGenR[36]	ShinyGO[35]
Publication year	**2022**	**2018**	**2020**	**2019**	**2021**	**2021**	**2020**
IMPORT DATA/MANAGEMENT:							
Count data input mode	x	x	x	x	x	x	
DESeq object input mode	x						
Gene list input mode	x						x
Several gene names	x	x	x	x	x	x	x
>2 species	x	x	x		x		x
Raw data accessibility	x	x	x	x	x		
Demo data	x	x	x	x			x
DEA COMPUTATION & DATA VISUALIZATION:							
DEA statistical calculation	x	x	x	x	x	x	
Manage stats threshold	x	x	x	x		x	
Interactive statistics summary table	x		x	x	x	x	
Interactive preview visuals	x	x	x	x	x	x	
Interactive DE visuals	x	x	x	x	x	x	
ENRICHMENT ANALYSIS & VISUALIZATION:							
Custom background universe	x						x
Split results by DE directions	x		x		x		
KEGG results	x	x		x	x	x	x
GO results	x	x	x	x	x	x	x
GSEA type results	x	x	x			x	
Interactive EA tables	x	x	x	x	x	x	x
Interactive EA visuals	x	x	x	x	x	x	x
DOWNLOAD & REPORT:							
Individual plots download (.SVG, .HTML, .PNG)	x	x	x	x	x	x	x
Interactive report	x		x	x			
Custom report	x						
ACCESSIBILITY:							
Public server	x	x	x	x			x
Source code available	x	x	x	x	x	x	x

Apart from these demo data, to facilitate the preparation of some input data for biologists who lack bioinformatics skills, we have provided on our GitHub two standard tutorials using the Galaxy platform.[34] The tutorials allow a user with no prior bioinformatics knowledge to generate read counts from raw RNA-seq .fastq data in two easily accessible step-by-step methods. Quality control and a few downstream analyses similar to those done within DEVEA can be explored following this pipeline. The tutorials can be accessed at https://github.com/MiriamRiquelmeP/DEVEA/blob/main/Galaxy_tutorials.md.

Related works (state-of-the-art)

The field of application tools for transcriptomics data visualization, DEA and EA is constantly growing. DEVEA functionalities are compared with six similar applications recently published (iDEP,[5] ShinyGO,[35] DEGenR,[36] GENAVi,[6] RNfuzzyApp[37] and ideal[7]). These tools operate through a graphical user interface, provide interactive results, and are based on stable and maintained R packages. A detailed comparison of the DEVEA main attributes is shown in Table 1. Characteristics that are related to data management and importation of data into the application have been stressed in the upper part of the table, followed by the various modes of DEG identification and different interactive graphical results, EA calculations and global reporting and webhosting. It is clear from Table 1 that most of the selected applications share many functionalities with DEVEA. One exception is the application ShinyGO, which offers significantly fewer options since it is designed only to perform enrichment calculations from simple gene lists. However, some other specific differences exist with other tools. DEVEA has more flexibility in terms of data type import in different formats, representing different stages of the analysis. Similarly, the user has a wide range of exploration possibilities via GL and GL + SV input formats, in terms of data generation and origin. The list of features and the associated statistics may have been generated from many different external tools or could represent several analysis types, as long as they are eventually converted into gene names (i.e. Microarray data results, proteomics results, gene lists from the literature, etc.). As a further advantage, the ability to import complex objects, such as DO, increases the number of visualization and analysis options. Despite this, not all possible visuals that can be generated from these objects are included in DEVEA, which has room for improvement. In particular, DEVEA could further develop data management functionalities, by extending the capacity of dealing with batch effects or data pre-processing and filtering. One of the possible drawbacks is the low number of available species for the EA compared with other tools or the potential mismatches when converting gene names. The EA, where there is information about the level of expression, can be generated from all DEGs, either split by the direction of expression change (only on up- or down-regulated genes), or merged into a single list. This is not always the case in applications that perform analyses on the whole list of potential genes of interest. The custom global report, a unique feature of DEVEA, might be very handy to share results with collaborators, because the user can easily insert comments and transfer the fully-interactive HTML report. Lastly, DEVEA appears slightly more flexible in terms of application hosting and running, since it is possible to run it online (DEVEA web server), or offline using R. For instance, certain applications such as DEGenR and RNfuzzyApp do not offer the possibility to run the application online.

Although overall applications share common analysis blocks, DEVEA presents more graphical variety than most of them. For example, as a criterion to consider in the table that an application contains “Interactive preview visuals” to preliminary explore the data, only one of PCA plot, violin/box plot of sample profile, heatmap for gene expression, sample clustering, gene expression dots plot per groups should be generated from the applications. For “Interactive DE visuals” of the DEA statistics and profile the criterion consist of including at least one of volcano plot, MA plot or karyoplot. Finally, for the “Interactive visuals” to navigate the EA, the marked tools include at least one plot among bars plot, dots plot, chord plot, heatmap, net plot, word cloud, circle plot of the enriched terms. For most applications, only a small subset of these plots are implemented. DEVEA contains all these graphs, requirement that none of the others met.

Conclusion and Future Perspectives

The DEVEA application is developed to improve the set of existing software to perform DEA, data visualization and annotation or EA from transcriptomics data. DEVEA meets the need for applications that give sufficient usage autonomy, without compromising the complexity and accuracy of the results. It provides an interactive and user-friendly interface accessible to users without bioinformatics training, with a high diversity of analysis components. Researchers can explore their data in real-time, carry out DEA and subsequent EA from distinct well-known databases without losing possible customization. DEVEA contains a large range of functions in a single tool, avoiding the use of different tools/websites to perform transcriptomics analysis, offering advanced ready-to-publish visuals, tables and results. One of the main strength is the incorporation of several input data types. Additionally, the use of robust R packages, especially the DEA DESeq2 package is an attractive functionality of the application. The possibility to include a custom background makes DEVEA better suited for analysis in which correction of some experimental bias could lead to better results in the EA, avoiding the inappropriate inflation of statistical p-values and false-positive results. Another key advantage is that DEVEA allows the user to extract their results individually or in an interactive format through a custom HTML format file. To further develop DEVEA analyses, we plan to offer additional pre-treatment options (e.g. removing batch effect, filtering genes

by expression, etc.) and integrate more species and include transcription factor enrichment analysis. In conclusion, the purpose of DEVEA is to promote the dialogue between biologists and bioinformaticians, particularly to produce suitable data and to understand the validity of the data needed to create the best downstream results.

Software availability

- Software available at http://shiny.imib.es/devea/. Archived source code as at time of publication: https://doi.org/10.5281/zenodo.6657245.
- Latest source code on https://github.com/MiriamRiquelmeP/DEVEA.
- Test files for every input mode can be found also on https://github.com/MiriamRiquelmeP/DEVEA/tree/main/data.
- Tutorial accessible from both DEVEA modules (DESeq DEVEA and Simple DEVEA) in the 'Tutorial' section from the top controls and independently on https://shiny.imib.es/DESeqDevea/tutorial.html or https://shiny.imib.es/simpleDevea/tutorial.html.

License: Apache license 2.0.

Author contributions

MRP and FPS conceived the application, did the development and wrote the manuscript. CE, SB and EB supervised the work, tested the application and wrote the manuscript. JFD contributed fund acquisition and resources for the project. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

We would like to thank all the people from MIRCen, CNRGH and IMIB who helped in the testing for the implementation of the application and provided valuable ideas for improvement. We would also like to thank Steven McGinn for his help in proofreading the manuscript.

References

1. Mortazavi A, Williams BA, McCue K, *et al.*: **Mapping and quantifying mammalian transcriptomes by rna-seq.** *Nat. Methods.* 2008; **5**(7): 621–628.
Publisher Full Text
2. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, *et al.*: **Translating rna sequencing into clinical diagnostics: opportunities and challenges.** *Nat. Rev. Genet.* 2016; **17**(5): 257–271.
PubMed Abstract | Publisher Full Text
3. R Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020.
Reference Source
4. Chang W, Cheng J, Allaire JJ, *et al.*: *shiny: Web Application Framework for R.* 2021. R package version 1.7.1.
Reference Source
5. Ge SX, Son EW, Yao R: **idep: an integrated web application for differential expression and pathway analysis of rna-seq data.** *BMC Bioinform.* 2018; **19**(1): 1–24.
6. Reyes ALP, Silva TC, Coetzee SG, *et al.*: **Genavi: a shiny web application for gene expression normalization, analysis and visualization.** *BMC Genomics.* 2019; **20**(1): 1–9.
7. Marini F, Linke J, Binder H: **ideal: an r/bioconductor package for interactive differential expression analysis.** *BMC Bioinform.* 2020; **21**(1): 1–16.
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): 1–21.
Publisher Full Text
9. Pagès H, Carlson M, Falcon S, *et al.*: *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor.* 2020. R package version 1.52.0.
Reference Source
10. Alexa A, Rahnenfuhrer J: *topGO: Enrichment Analysis for Gene Ontology.* 2020. R package version 2.42.0.
11. Korotkevich G, Sukhov V, Sergushichev A: **Fast gene set enrichment analysis.** *bioRxiv.* 2019.
Publisher Full Text | Reference Source
12. Guangchuang Y, Wang L-G: **Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters.** *Omics: A Journal of Integrative Biology.* 2012; **16**(5): 284–287.
13. Wickham H: *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag; 2016.
Reference Source
14. Sievert C: *Interactive Web-Based Data Visualization with R, plotly, and shiny.* Chapman and Hall/CRC; 2020.
Reference Source
15. Baik B, Yoon S, Nam D: **Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data.** *PLoS One.* 2020.
Publisher Full Text
16. Ritchie ME, Phipson B, Wu D, *et al.*: ***limma* powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; **43**(7): e47.
Publisher Full Text
17. Hardcastle TJ: baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. *R package version 2.31.0.* 2022.
18. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
Publisher Full Text
19. Oleś A: **Deformats: Differential gene expression data formats converter.** 2021. R package version 1.20.0.
Reference Source

20. Langley SR, Mayr M: **Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics.** *J. Proteome.* 2015; **129**: 83–92.
PubMed Abstract | Publisher Full Text

21. Zhang Y, Parmigiani G, Johnson WE: **ComBat-seq: batch effect adjustment for RNA-seq count data.** *NAR Genom. Bioinform.* 2020; **2**(3): lqaa078.
Publisher Full Text

22. Altman DG, Bland JM: **Measurement in medicine: The analysis of method comparison studies.** *Journal of the Royal Statistical Society. Series D (The Statistician).* 1983; **32**(3): 307–317. 00390526, 14679884.
Reference Source

23. Kanehisa M, Goto S: **Kegg: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Botstein D, Cherry JM, Ashburner M, *et al.*: **Gene ontology: tool for the unification of biology.** *Nat. Genet.* 2000; **25**(1): 25–29.

25. Jiang Z, Gentleman R: **Extensions to gene set enrichment.** *Bioinformatics.* 2007; **23**(3): 306–313.
PubMed Abstract | Publisher Full Text

26. Abjean L, Ben Haim L, Riquelme-Perez M, *et al.*: **Reactive astrocytes promote proteostasis in Huntington's disease through the JAK2-STAT3 pathway.** *Brain.* 03 2022; awac068.
Publisher Full Text

27. Tabrizi SJ, Flower MD, Ross CA, *et al.*: **Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities.** *Nat. Rev. Neurol.* 2020; **16**(10): 529–546.
PubMed Abstract | Publisher Full Text

28. Verkhratsky A, Nedergaard M: **Physiology of astroglia.** *Physiol. Rev.* 2018; **98**(1): 239–389.
PubMed Abstract | Publisher Full Text

29. Ben Haim L, Sauvage M-A C-d, Ceyzériat K, *et al.*: **Elusive roles for reactive astrocytes in neurodegenerative diseases.** *Front. Cell. Neurosci.* 2015; **9**: 278.
Publisher Full Text

30. Andrews S: *FastQC: A Quality Control Tool for High Throughput Sequence Data.* 2010.
Reference Source

31. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat. Methods.* 2015; **12**(357-360): 9.

32. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**: 923–930.

33. Lee K, Lehmann M, Paul MV, *et al.*: **lack of FIBRILLIN6 in Arabidopsis thaliana affects light acclimation and sulfate metabolism.** *New Phytol.* 2020; **225**(4): 1715–1731.

34. Afgan E, Baker D, Batut B, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses.** *Nucleic Acids Res.* 2 July 2018; **46**(W1): W537–W544.
Publisher Full Text

35. Ge SX, Jung D, Yao R: **Shinygo: a graphical gene-set enrichment tool for animals and plants.** *Bioinformatics.* 2020; **36**(8): 2628–2629.
PubMed Abstract | Publisher Full Text

36. Choudhary KS, Caldwell AB, Subramaniam S: **DEGenR: An R Shiny app for differential gene expression and enrichment analysis.** May 2021.
Publisher Full Text

37. Haering M, Habermann BH: **Rnfuzzyapp: an r shiny rna-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis.** *F1000Res.* 2021; **10**: 654.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 27 March 2023

https://doi.org/10.5256/f1000research.144696.r167705

✔ **Wenbin Guo**
Information and Computational Sciences, James Hutton Institute, Dundee, UK

In response to the feedback provided in the last review, the authors have carefully reviewed and revised their manuscript, taking into account all the comments and suggestions made. They have added new information where necessary to further support their claims. The authors have made sure that their arguments and findings are presented clearly and logically. Overall, the authors have shown great attention to detail and have made improvements to their manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics method development, automated pipeline development, Shiny App development, bioinformatics analysis, RNA-seq data analysis, differential gene expression analysis, differential alternative splicing analysis, gene regulatory network analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 October 2022

https://doi.org/10.5256/f1000research.135003.r152235

Wenbin Guo
[1] Information and Computational Sciences, James Hutton Institute, Dundee, UK
[2] Information and Computational Sciences, James Hutton Institute, Dundee, UK

Summary

The author developed a shiny App DEVEA for biologists who lack extensive bioinformatics skills to perform complex differential expression analysis and downstream function annotation and pathway enrichment analysis. It provides an easy-to-use platform and empowers biologists to hand on complex RNA-seq analysis by themselves. The multiple options of inputs, including (a) the raw read counts of genes, (b) the intermediate dataset produced by DESeq2, (c) the gene list with statistics of significance and (d) the gene list only, allowed users to start the analysis flexibly from different types of data. It also provided interactive figures and tables to visualise the results in real-time. In the end, a report can be generated with user-selected results and figures, which improved the reproducibility of the RNA-seq analysis.

Below are some limitations of the current DEVEA and suggestions for future improvements. The author already had a plan to address some of them in future directions.

- It seems that the current version of DEVEA only supports the analysis of human, mouse and rat, which limits its usage in wider studies of plant species and other animal species.

- Due to the above limitation, the current manuscript lacks some case studies in other species, especially in plants.

- The controls for errors caused by unwanted variations are not proposed in the DEVEA pipeline. For example, RNA-seq data often have batch effects between replicates and they will largely hinder the accuracy of differential expression analysis. DESeq2 also provides suggestions on how to correct the batch effects in the user manual.

- DEVEA only allows a pair-wise comparison between control and treatment at a time, which is not flexible for complex experimental design, such as datasets with a great number of conditions to compare, datasets of time-series data and development series, and datasets with multiple groups and each group include addition level of conditions to compare.

- It would be impressive to enable enrichment analysis for the genes lacking annotations. For example, getting the gene annotation by blasting the sequence against some databases. It will be especially useful for the analysis of newly assembled genes and transcripts.

Comments for manuscript improvement:

- The preparation of some input data is still difficult for biologists who lack bioinformatics skills. For example, getting the raw read counts requires bioinformatics skills to map the RNA-seq data to a reference genome or transcriptome and generate read counts by using quantification tools such as Salmon, Kallisto and RSEM. The author should propose an easy-to-access method and provide step-by-step tutorials for biologists to generate the read counts from raw RNA-seq data, for example, using the Galaxy platform.

- There are massive methods available for the function of differential expression (DE) analysis. Many comparison studies showed that the methods DESeq2, Limma-voom, edgeR

and Sleuth have similar performance in DE analysis and each of these methods has a big user group. To attract more users to DEVEA, the author could highlight the advantages to use DESeq2 in DEVEA compared with other methods, such as Limma and edgeR, in the introduction or discussion.

- In the tutorial, the table of contents is floating and overlaps with texts and videos. Is it possible to reallocate the position of the table of contents to avoid overlapping?

- In the example data link: https://github.com/MiriamRiquelmeP/DEVEA/tree/main/data, it would be better to include a readme file with descriptions of the example datasets.

- "Supplemental-Information.txt: Running DEVEA locally on a machine": I would recommend the author build DEVEA as an R package. Then, instead of "Download the compressed folder from the DEVEA GitHub repository", users can install and run DEVEA on a local machine as an R package.

- Page 3: "Despite the great progress to facilitate transcriptomics computational analyses, some improvements are still possible for these tools." Although comparisons were given in the Related work section, the author should give brief descriptions of these tools and summarise their limitations as literature reviews in the introduction. The author also should address what are the possible improvements and why they are needed.

- Page 11: "Reads were aligned on the mouse genome (ENSEMBL GRCm38 release 96) and a count matrix was generated." For reproducibility, the tool for read mapping and count matrix generation should be referred to and cited.

- Figure 7: the font size of labels and some details of the figures are too small. Would it be better to change the multiple plot layout to 3 rows x 2 columns?

The paper has quite a few typos. The author should carefully check the entire manuscript again. Here, I have picked out some of them:

- Page 3: "key steps of a complete RNA-seq analyses pipeline" should be "key steps of a complete RNA-seq analysis pipeline"

- Page 3: "detailed bellow in the circular notes" should be "detailed below in the circular notes".

- Page 5: "the simple gene list (GL) can be build from" should be "the simple gene list (GL) can be built from"

- Page 6: "the program extract the conditions and calculates" should be "the program extracts the conditions and calculates"

- Page 11: "To demonstrate the usefulness of DEVEA, a RNA-seq" should be "To demonstrate the usefulness of DEVEA, an RNA-seq"

- Page 12: "which offers significantly less options since" should be "which offers significantly fewer options since"

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics method development, automated pipeline development, Shiny App development, bioinformatics analysis, RNA-seq data analysis, differential gene expression analysis, differential alternative splicing analysis, gene regulatory network analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 Feb 2023

Miriam Riquelme-Perez

We thank the reviewer for taking the time to review our paper and test our application. We provide here a detailed response (in bold, after + symbol) to all the comments (introduced by -, in italic) and how we addressed the majority of them.

- It seems that the current version of DEVEA only supports the analysis of human, mouse and rat, which limits its usage in wider studies of plant species and other animal species. Due to the above limitation, the current manuscript lacks some case studies in other species, especially in plants.
+ We acknowledge that there is limited support for plant and animal species in the current version of DEVEA. We are planning to expand the number of species available in future versions. For this version, following the reviewer's suggestion, we have now included the model plant species *Arabidopsis thaliana* in DEVEA. A demo dataset has also been created compatible with DEVEA for this species, based on an RNA-seq study published recently (Lee et al. New Phytologist, 2020, 225(4):1715-1731).

- The controls for errors caused by unwanted variations are not proposed in the DEVEA pipeline. For example, RNA-seq data often have batch effects between replicates and they will largely hinder the accuracy of differential expression analysis. DESeq2 also provides suggestions on how to correct the batch effects in the user manual.

+ We can distinguish between modeling the batch effect using DESeq2 within DEVEA and removing the batch effect before importing the data into DEVEA. An advantage of using as input data a DO generated previously in R, is that the user can include Batch in the design formula as "design = ~Batch + Condition". Once the final contrast is specified as Condition_level1_vs_level2, the batch effect will then be 'accounted for' (the statistical inferences will be adjusted for Batch) with a differential expression analysis testing for Condition. Batch is essentially treated in DEVEA as a covariate in the regression model, just as we do in association studies. If the user wants to effectively remove the batch effect for downstream analyses, it can be performed outside DEVEA on the variance-stabilized or rlog expression values using limma::removeBatchEffect() or ComBat-seq tool, before uploading the data to the application again. This could be considered for future versions of DEVEA as well. We have added a section in the 'Data upload and statistical design specification' section of the manuscript.

- DEVEA only allows a pair-wise comparison between control and treatment at a time, which is not flexible for complex experimental design, such as datasets with a great number of conditions to compare, datasets of time-series data and development series, and datasets with multiple groups and each group include addition level of conditions to compare.

+ Multiple experimental conditions can be included in the design of the DESeq2 object (DO). DEVEA will offer the possibility to consider any of them foreach analysis. For instance, when "design = ~ConditionA + ConditionB + ConditionC", the tool will propose as possible contrast 1.ConditionA_level1_vs_level2, 2.ConditionB_level1_vs_level2 and 3.ConditionC_level1_vs_level2. The user can select from DEVEA interface, which specific contrast to explore. In addition, for DO and CM + SI, when there are more than two levels from the same condition, DEVEA will propose all the one-vs-one combinations. Therefore, if the "design = ~ ConditionWith4levels", the possible contrasts would be 1.Condition_level1_vs_level4, 2.Condition_level2_vs_level4, and 3.Condition_level3_vs_level4, level4 being the basal condition for every contrast. For CM + SI, these levels are selected alphabetically. In a DO, the user can use the function relevel() from DESeq2 in R to specify the basal level. We have changed the text in the manuscript to indicate these possibilities, in the 'Data upload and statistical design specification' from the 'Implementation' section.

- It would be impressive to enable enrichment analysis for the genes lacking annotations. For example, getting the gene annotation by blasting the sequence against some databases. It will be especially useful for the analysis of newly assembled genes and transcripts.

+ We agree that such an initiative would provide a genuinely useful service, especially for recently sequenced organisms who do not have yet a good and comprehensive annotation. However, this task is far from being trivial and requires a substantial amount of software development and is therefore currently beyond the scope of the current DEVEA implementation. For DEVEA, we currently aim to exploit available

packages and resources for visualization, differential expression analysis and enrichment from well-known and described species.

Comments for manuscript improvement:
- The preparation of some input data is still difficult for biologists who lack bioinformatics skills. For example, getting the raw read counts requires bioinformatics skills to map the RNA-seq data to a reference genome or transcriptome and generate read counts by using quantification tools such as Salmon, Kallisto and RSEM. The author should propose an easy-to-access method and provide step-by-step tutorials for biologists to generate the read counts from raw RNA-seq data, for example, using the Galaxy platform.
+ It is true that data preparation can be quite complicated for non-specialists. Following the remark of the reviewer, we have added references and web links for two complete pipelines and tutorials for Galaxy on our DEVEA GitHub page (https://github.com/MiriamRiquelmeP/DEVEA/blob/main/Galaxy_tutorials.md). We have also included a brief explanation of them in the 'Use case' section.

- There are massive methods available for the function of differential expression (DE) analysis. Many comparison studies showed that the methods DESeq2, Limma-voom, edgeR and Sleuth have similar performance in DE analysis and each of these methods has a big user group. To attract more users to DEVEA, the author could highlight the advantages to use DESeq2 in DEVEA compared with other methods, such as Limma and edgeR, in the introduction or discussion.
+ It is correct that there are several methods available for RNA-seq differential expression analysis with comparable performances. We chose DESeq2 because it is widely used, very robust and recognised in the community. As suggested by the reviewer, we now stress, in the 'Operation' and briefly in the 'Conclusion and Future Perspectives' sections, the robustness of the method by mentioning an article that compares DESeq2's performance with other methods (Baik et al., PLoS One, 2020, 15(4):e0232271).

- In the tutorial, the table of contents is floating and overlaps with texts and videos. Is it possible to reallocate the position of the table of contents to avoid overlapping?
+ We have noticed that the HTML file that displays the tutorial is significantly heavy. As a result, some browsers do not render it properly and take some time to put each element in its place. In order to view the tutorial, as well as the complete workflow of DEVEA itself, we strongly recommend the use of Mozilla Firefox Web Browser. This part has been clarified in the 'Operation' section.

- In the example data link: https://github.com/MiriamRiquelmeP/DEVEA/tree/main/data, it would be better to include a readme file with descriptions of the example datasets.
+ A README.md file has been included on GitHub data folder, with all the test data inside (for Mus musculus and for Arabidopsis thaliana).

- "Supplemental-Information.txt: Running DEVEA locally on a machine": I would recommend the author build DEVEA as an R package. Then, instead of "Download the compressed folder from the DEVEA GitHub repository", users can install and run DEVEA on a local machine as an R package.
+ In fact, DEVEA is a web interface with two R Shiny applications bundled together and due to R packages structural constraints, it is not so straightforward to package such

an application. However, to facilitate the local installation of DEVEA, we have added a script on DEVEA's github that will install all the necessary R dependences for DEVEA (script_libraryApp.R). Then the user only has to clone the DEVEA repository from github and launch the application (2-steps process) after running the script. For future versions, we plan to develop a Docker image, rather than an R package. We have detailed the process in the supplemental information (https://github.com/MiriamRiquelmeP/DEVEA/blob/main/Supplemental-Information.txt)

- Page 3: "Despite the great progress to facilitate transcriptomics computational analyses, some improvements are still possible for these tools." Although comparisons were given in the Related work section, the author should give brief descriptions of these tools and summarise their limitations as literature reviews in the introduction. The author also should address what are the possible improvements and why they are needed.
+ We have added a short paragraph in the introduction summarizing the features and limitations of the tools comparable to DEVEA in the introduction.

- Page 11: "Reads were aligned on the mouse genome (ENSEMBL GRCm38 release 96) and a count matrix was generated." For reproducibility, the tool for read mapping and count matrix generation should be referred to and cited.
+ The software tools we used for this task and further details are now indicated in the main text.

- Figure 7: the font size of labels and some details of the figures are too small. Would it be better to change the multiple plot layout to 3 rows x 2 columns?
+ We have tried to reformat the figure layout to 3 rows x 2 columns, but the result was really not satisfying, so we preferred to keep the original layout.

The paper has quite a few typos. The author should carefully check the entire manuscript again. Here, I have picked out some of them:
Page 3: "key steps of a complete RNA-seq analyses pipeline" should be "key steps of a complete RNA-seq analysis pipeline"

Page 3: "detailed bellow in the circular notes" should be "detailed below in the circular notes".

Page 5: "the simple gene list (GL) can be build from" should be "the simple gene list (GL) can be built from"

Page 6: "the program extract the conditions and calculates" should be "the program extracts the conditions and calculates"

Page 11: "To demonstrate the usefulness of DEVEA, a RNA-seq" should be "To demonstrate the usefulness of DEVEA, an RNA-seq"

Page 12: "which offers significantly less options since" should be "which offers significantly fewer options since"
+ We thank the reviewer for pointing out those typos. All of them have been corrected.

We have also carefully proofread the manuscript for other remaining typos.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 August 2022

https://doi.org/10.5256/f1000research.135003.r142536

Hélène Hirbec
[1] Institut de Génomique Fonctionnelle, CNRS, INSERM, Université de Montpellier, Montpellier, France
[2] Institut de Génomique Fonctionnelle, CNRS, INSERM, Université de Montpellier, Montpellier, France

In the present manuscript Riquelme-Perez and coll. are presenting the DEVEA software aimed at helping biologists with no or poor programming skills to analyse their Omics data (mainly RNA-seq). The rationale for the development of a new software is clearly explained, additionally in the manuscript, the authors compare their software to other web-based software and highlight that DEVEA is higher flexibility (see Table 1). The software is available as a web-based application but can also be installed locally.

The description of the software is technically sound. As a matter of facts, it is based on the several well-known and highly used R packages aimed at analysing RNA-seq data. Interestingly, differential expression analysis (DEA) is performed using DESeq2 package on the most reliable tool for DEA. DEVEA is also flexible as it accepts input data from different format and at different stage of the analysis. Thus, DEVEA can be used for initial analysis of the data, but can be used at later stage of the data analysis. One very interesting feature of the DEVEA software is that it enables loading a specific background universe.

The manuscript is clear and very well written. There is sufficient information provided in the manuscript to run the analyses proposed. However, it will be useful if the authors could provide (as supp material?) a step by step protocol for the different type of analysis (CM+SI, DO, GK+SV; GL).

Overall, the manuscript presents a new software that appear more flexible and present several improvements compared to other similar software developed recently. DEVEA will be useful to biologists with poor programming skills to explore their omics data, in particular RNA-seq data. Although, the software is already usable, I noticed a couple of minor issues that would deserve improvements.

Minor points that would need improvements:

- Could the authors specify if already normalized counts can be used as input in CM+SI type of analysis?
- Could the authors specify whether the software support both "." and "," as decimal separators;
- When uploading the GL with gene symbols, these latter are automatically converted to Ensembl identifiers, but some symbol are not recognized. It is unclear which genes are considered for EA analyses: all genes with symbol, only genes with an Ensembl identifier, only annotated genes with an Ensemble identifier?
- In some of the EA graphs, deregulated pathways/gene sets are referred to by their url rather than by the pathway name. This complicates the reading of the graphs;
- Demo data are downloadable, but the link given in the manuscript does not work;
- Downloads of the SVG images don't seem to work, except for Chordplot in KEEG Enrichment for which the download is different. However, in this latter case, only the plot (but not the legend) is downloaded;
- The inclusion of GSEA type results in the analysis is interesting. Including the GSEA leading edge analysis in DEVEA would be an interesting add.

Further improvements that would be interesting to make:

- Enrichment analysis: In addition to GO and KEEG analyses, it would also be interesting to add other databases, such as Wiki Pathways for example.
- It will also be interesting to enable evaluating over-representation for custom genesets (i.e. specific molecular signatures of interest in the research field).

Remarks: Assessment as to whether sufficient details of the code, methods and analysis is provided to allow replication of the software development and its use by others, is beyond the competence of the reviewer.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics; Neuroinflammation; Alzheimer Disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 20 Feb 2023

Miriam Riquelme-Perez

Thank you for taking the time to review our paper, and for the thoughtful suggestions. In the revised version, we have modified the main text to refer more clearly to the method information provided on the interface, and improved the clarity of the text in several places, according to your suggestions. In the same way, we have included a few enhancements in the application tool itself.

Below are the questions raised by the reviewer (- in italic), followed by the authors' response (+ in bold):

- The manuscript is clear and very well written. There is sufficient information provided in the manuscript to run the analyses proposed. However, it will be useful if the authors could provide (as supp material?) a step by step protocol for the different type of analysis (CM+SI, DO, GK+SV; GL).
+ In addition to the article and the information within DEVEA, it was also possible to consult a comprehensive walkthrough on how to use the application from the different input modes. It was accessible from both DEVEA modules (DESeq DEVEA and Simple DEVEA) in the 'Tutorial' section from the top right controls inside the application interface and independently on https://shiny.imib.es/DESeqDevea/tutorial.html or https://shiny.imib.es/simpleDevea/tutorial.html. The tutorial was mentioned in the 'Software availability' section, and now this is mentioned in the 'Operation' section.

- Could the authors specify if already normalized counts can be used as input in CM+SI type of analysis?
+ The counting matrix (CM) and the sample information (SI) elements will be used as raw data to be internally analyzed through the various functions of the DESeq2 package. To build the final DESeqDataSet object, DEVEA uses the data as obtained from the counting process, presented in the form of an array of integer values. Decimals will not be accepted. The DESeq2 model internally corrects for library size and generates its own normalization, so transformed or normalized values must not be used in the CM + SI input type (as detailed in the official bioconductor DESeq2

vignette section "Why un-normalized counts?" http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#why-un-normalized-counts). A sentence that specifies this point is now included in the 'Data requirement' section, 1. CM + SI mode.

- Could the authors specify whether the software support both "." and "," as decimal separators?
+ Thank you for raising this point. Decimals must be indicated with "." as a separator. "," will not be properly recognized by the program.

- When uploading the GL with gene symbols, these latter are automatically converted to Ensembl identifiers, but some symbol are not recognized. It is unclear which genes are considered for EA analyses: all genes with symbol, only genes with an Ensembl identifier, only annotated genes with an Ensemble identifier?
+ The KEGG, GO, and GSEA libraries used in DEVEA need as input a list of genes coded in their ENTREZ ID format. For this reason, an exhaustive process to convert the initial gene names provided by the user (either in gene SYMBOL or ENSEMBL) into ENTREZ ID is conducted internally by DEVEA. Finally, the EA functions retrieve the associated annotation terms from these genes to generate the results. Therefore, only genes from the original data with an ENTREZ ID name can be linked to its annotation in the libraries.
Unfortunately, not all genes have curated ENTREZ ID names linked with the different formats or have annotated functions associated with the libraries. As a consequence, some genes will be left out of the EA. The number of genes considered in the following steps is indicated in green below the "preview table" as shown in Figure 5B. This is a limitation of the gene name conversion functions without manual curation. This point was discussed in the 'Enrichment analysis (EA) and visualization' subsection in the 'Implementation' part.

- In some of the EA graphs, deregulated pathways/gene sets are referred to by their url rather than by the pathway name. This complicates the reading of the graphs;
+ We agree that using only the URL is cumbersome since it is necessary to go back to the table to link with the actual name. Due to space limitations in some plots, pathways or functions that have a very long name cannot be shown in full. However, in most of the images where the full name cannot be displayed, it will be shown in full by placing the mouse arrowhead on top of the interactive graph.

- Demo data are downloadable, but the link given in the manuscript does not work;
+ We apologize for this problem. The link has been corrected in the manuscript and now redirects to the appropriate page to access the demo data on GitHub.

- Downloads of the SVG images don't seem to work, except for Chordplot in KEGG Enrichment for which the download is different. However, in this latter case, only the plot (but not the legend) is downloaded;
+ Indeed, there was a server issue that has now been fixed. The functionality to export SVG images is fully operational again. Some graphics may change slightly when downloaded because the functions used to render/download are not necessarily the same as those used to display them. The plot information must remain the same.

- The inclusion of GSEA type results in the analysis is interesting. Including the GSEA leading edge analysis in DEVEA would be an interesting add.
+ We agree that the inclusion of the leading edge for GSEA results is interesting. This has been included in DEVEA as an extra column in the table at the GSEA tab, called "leading_edge". It has also been implemented in the plot as a red line indicating the extent of what we consider the leading edge genes. This is also mentioned in the text 'Enrichment analysis (EA) and visualization' section.

- Further improvements that would be interesting to make :
- *Enrichment analysis: In addition to GO and KEEG analyses, it would also be interesting to add other databases, such as Wiki Pathways for example.*
- *It will also be interesting to enable evaluating over-representation for custom genesets (i.e. specific molecular signatures of interest in the research field).*

+ Regarding the first further improvement, adding more databases would be a significant extension to the current version of DEVEA, that we want to keep fast and operational. For the second suggestion, the specific custom genesets needed for the over-representation analysis could be very different between users and experiments, hence we prefer to only include the curated well known genesets from the databases. Nevertheless, DEVEA allows the definition of background genes for comparison in the EA part, avoiding the inappropriate inflation of statistical p-values and false-positive results from contrast against the whole genome/proteome.
For the moment, we aim at offering a stable and robust version and creating a docker container with these basic functionalities efficiently working in the server. We think that including these significant improvements is beyond the scope of this first DEVEA version. However, these are exciting suggestions to be considered in the following versions. Our goal is to further develop DEVEA according to the needs of the community and bio-informatics developments. This is acknowledged in the 'Conclusion and Future Perspectives' section, and we will consider including new functionalities in a future version of DEVEA.

Competing Interests: No competing interests were disclosed.