# DESeqUI

This shiny app is based on the package DESeq2 and provides a user interface for RNAseq data analysis. The app is designed for working with **Trypanosoma brucei** RNAseq data. Feel free to fork this repo and modify it as you wish. Pull requests are welcome.

## Dependencies

- shiny
- DESeq2
- ggplot2
- plotly

## How to use DESeqUI

### Setup

It is recommended to use RStudio.

- Download the package
- Unzip the package
- open DESeqUI/app.r in Rstudio
- Click 'Run App' (top right corner)

The program will load its dependencies into the environment (observe RStudios console) and the user interface will appear. On the left, there is the settings panel (see screenshot) which allows the user to define specific parameters for analysis. At the top of the app window, the tab-bar is located. Each tab shows the results of another analysis method. From left to right: PCA (shows a principal component analysis of the data), DESeq (MA-plot after DESeq2 analysis), Classes (class enrichment analysis), Results Table (Shows the final table for review and download).

- Click 'Browse' (top left in the app) and select the Count-table (tab or comma separated, first column geneIDs, following columns raw read counts of individual samples.)
- Below the 'Browse'-Button, details about the file can be specified. If you used TrypRNAseq for data preparation, the default can be used
- If the upload of the data worked properly, the first couple of line of the file will be shown in the main-window
- If no preview appears in the 'Content'-tab, the data was not loaded correctly. Please make sure to specify the correct parameters on the left, according to your file
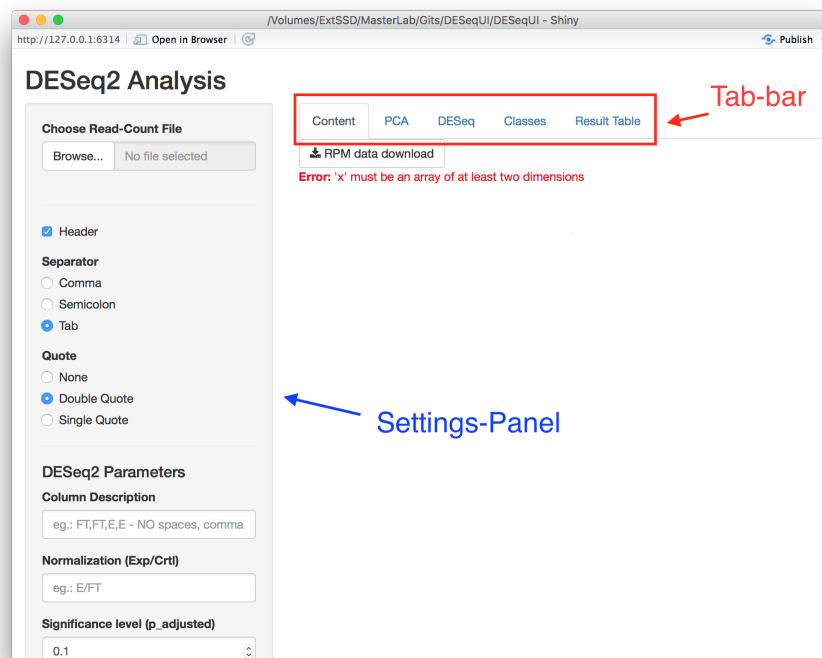
Figure 1: General app layout.

- You can download the reads RPM (reads per million) transformed (RPM data download button) or as regularised logarithm (rlog) transformed data (Rlog data download button). Before the latter can be used, please specify the DESeq2 parameters and click apply at the buttom of the settings panel



Figure 2: Parameters for correct file input.

Now set up the DESeq2 parameters:

**Column description**: Group the columns in the read file in their corresponding groups. For example, if the first three samples (first three columns) are treated and the following two control, write: treat,treat,treat,ctrl,ctrl

The descriptions can be anything, but be consistent. You could have also written: t,t,t,c,c

Please make sure not to use spacebar, just write down a comma-seperated list.

**Normalization**: Here, the contrast which should be studied can be defined. To keep our previous example, we are interested in significant differences between the treated and the control group. Therefore, we should enter: treat/ctrl

If the data-frame consists of multiple groups, pick the ones you want to study.

**Significance level**: Here you can define which level of significance you want to set. The default DESeq2 value is 0.1 which was kept in DESeqUI.

**Alternative Hypothesis**: As the name implies, this option defines the alternative hypothesis which should be tested for, depending on the experimental design: If you compare full transcriptomes and you want to see the full effect of treatment against control, keep the option at 'both'. If you did a pulldown and you only want to see, what is interacting with your protein, set it to 'Greater'.

The class enrichment parameters can be left as they are for now.

**Data analysis**

- Click on the PCA tab In the RStudio console you should see the programm working. Depending on the data-size, this can take some minutes. When the calculations are done a PCA-plot (principle component analysis) will be shown. The plot is interactive and by hovering over each point, its identity is shown as a tooltip

- Click on the DESeq tab A MA-plot will be generated according to the DESeq2 package description. For details please check out the vignette of DESeq2. In short: Each point represents one gene. Points which are colored red were identified as significantly (see above) different in both specified groups.

- Click on the Classes tab Here, class enrichment of significantly differential expressed genes (according to DESeq2) analysis can be done. Classes are specified using a curated list of genes (Prof. Christine Clayton). Using the **Class enrichment Parameters** panel, you can setup the analysis. If you are interested in the classes of genes which are upregulated (at least 2-fold) set up as follows: log2FoldChange -> greater, Cutoff log2FoldChange -> 1. Now, the program counts the class appearance in the defined subset of genes and tests whether each class appears more often than in the background (the unique gene list), using Fishers exact test and adjustment of the p-value using the Benjamini-Hochberg method. The uppermost barplot shows the result of this analysis. Only classes below the user-defined level of significance will be reported. The p-adjusted value is displayed below each bar-group. If a 0 is printed, the p-adjusted value is below 0.001. The exact value is printed to the RStudio console.

The next plot shows a boxplot which reports the distribution of each class within **all** significantly different genes in both groups. The red line indicates the log2FoldCutoff defined previously.

The third plot, another barplot, reports the peak time during cellcycle for each subset significant gene. This plot is affected by the log2FoldChange option and the cutoff.

4

The scatterplot below resembles in its coloring the MA-plot but shows the transcript length for each gene against their log2Fold change. Here, a length-bias could be identified. For example you did a RITseq experiment and the matrix you were using interacts with RNA itself. Therefore the longer the RNA the more it binds to the matrix. A bias like that would be visible in this plot.

**Attention**: If you specified in the DESeq2 parameters an alternative hypothesis of 'Greater' you can not do the class enrichment analysis for genes which show fewer mRNA abundance, obviously. If you want to do the class enrichment analysis on full transcriptomes, so that you can look at the up AND downregulated genes, keep the DESeq2 parameter 'Alternative Hypothesis' at 'both'!

**Attention 2**: The 'Class enrichment parameter' 'log2Fold Change' specifies the subset of genes which show either higher ('Greater') or lower ('Less') abundance than control. The option 'both' is only added to have the boxplot across the whole range of significant (according to DESeq2) changes. The first plots output, if 'log2Fold Change' set to 'both', does not give any result, obviously.

- Click on Result Table This is a preview of the DESeq2 output table. It is subset using the Class Enrichment Parameters defined by the user. The full table can be downloaded using the download button.

## Trouble shooting

### 'Content'-tab: 'x' must be an array of at least two dimensions'

At the start of the app, the following error will be shown in the 'Content'-tab: ''x' must be an array of at least two dimensions'. This is normal behaviour, since no data-frame was inputed. Upload a data frame using the 'Browse' Button.

### 'PCA'-tab: arguments imply differing number of rows: n, m

This error occurs if the defined 'Column description' (DESeq2 parameters, settings panel) contains another number of instances than the number of culumns of the provided data-set. Make sure to define for each column the right group.

### 'Classes'-tab: Aesthetics must be either length 1 or the same as the data (1): y, label, x, fill

If this error occurs where normaly a barplot was expected (at the top of the 'classes'-tab), it will be shown, since there is no class significantly over-represented. Try to relax your restrictions in the 'Class enrichment Parameters' in the settings panel.

# References

Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550. doi: 10.1186/s13059-014-0550-8