Instructions for the RSDV App

The RSDV app allows users to visualize differentitally expressed genes(DEG) starting with count data.

- Explore the app's features with the example data set pre-loaded by clicking on the tabs above.
- Upload your data manually

Information for the RSDV App

The app is hosted on shinyapps: https://kcvi.shinyapps.io/shiny-RSDV/
Code can be found on github: https://github.com/starryHK/shiny-RSDV
To run this app locally on your machine, download R or RStudio and run the following commands once to set up the environment:

```
install.packages(c("shiny","ggplot2","gplot2",DESeq2","RColorBrewer","DT"))
source("https://bioconductor.org/biocLite.R")
biocLite(c("voom","edgeR"))
```

You may now run the shiny app with just one command in R:

```
shiny::runGitHub("RSDV")
```

Input Data

You may use this app by

- Exploring the pre-loaded example data set. This is a pre-loaded mouse macrophages

RNA-seq example for exploring the app's features.

- Upload your own data that is Count data (or log2-expression data) which come from transcriptome sequencing.

Data Format

- File must be the row counts, not normalized data, e.g. FPKM, TPKM,TPM.
- File must have a header row.
- First/Left-hand column(s) must be gene identifiers.
- First/Left-hand column(s) must be defined as row_name.

Example of Data format

Each row denotes a gene, each column denotes a sample.

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

Analysis: When raw counts are uploaded, the data is then analyzed by the app. The app uses the voom method from the 'limma' Bioconductor package to transform the raw counts into logged and normalized intensity values. These values are then analyzed via linear regression where gene intensity is regressed on the group factor. P-values from all pairwise regression tests for group effect are computed and Benjamini-Hochberg false discovery rate adjusted pvalues are computed for each pairwise comparison.

DEG Table

- Column A provide gene name.
- Column C and column G provide Fold Changes and FDR,respectively.
- We use both log2FC and FDR to file DEG.

	A	B	C	D	E	F	G
1	gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
2	ENSMUSG00000079297	64.3235288	-9.693603725	1.2435323	-7.795217	6.43E-15	6.85E-14
3	ENSMUSG00000032401	319.104289	-5.358827278	0.2708604	-19.78446	4.05E-87	1.20E-83
4	ENSMUSG00000021250	16938.3171	-5.269720043	0.249842	-21.09221	9.38E-99	1.11E-94
5	ENSMUSG00000003545	4117.28088	-4.760813122	0.2763741	-17.22598	1.70E-66	2.51E-63
6	ENSMUSG00000043953	100.969013	-4.685822405	0.8536473	-5.489179	4.04E-08	1.66E-07
7	ENSMUSG00000022997	24.7484764	-4.595420405	0.7214419	-6.369772	1.89E-10	1.09E-09
8	ENSMUSG00000032218	158.158731	-4.369895934	0.3666836	-11.91735	9.61E-33	9.61E-31
9	ENSMUSG00000044244	133.023951	-4.229437558	0.3767673	-11.2256	3.05E-29	2.01E-27
10	ENSMUSG00000114277	89.3273473	-4.186160219	0.3934091	-10.64073	1.93E-26	8.89E-25
11	ENSMUSG00000055148	4483.51471	-4.162121229	0.2099188	-19.82729	1.73E-87	6.82E-84
12	ENSMUSG00000059743	5079.14998	-4.107194978	0.2647328	-15.51449	2.77E-54	1.82E-51

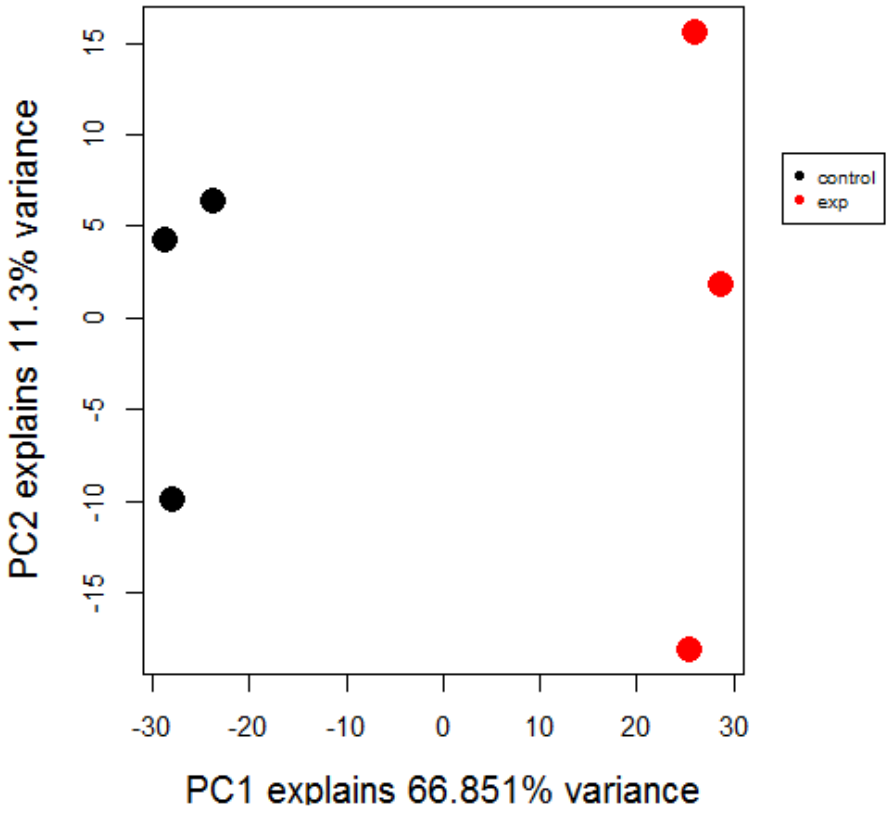
Analyzed data must contain some kind of expression measure for each sample (i.e. counts, normalized intensities, CPMs), and a set of p-values with corresponding fold changes for those p-values. For instance, if you have a p-value for the comparison of control vs exp , you can upload the observed fold change or log2(fold change) between control vs exp. If you have a more complex design and do not have fold changes readily available, you may upload the test statistics or other similar measures of effect size as placeholders. The fold changes are mainly used in the volcano plots. We recommend uploading p-values that are adjusted for multiple comparisons (such as q-values from the qvalue package, or adjusted p-values from p.adjust() function in R).

Visualization

Group Plots

PCA Plot

This plot uses Principal Component Analysis (PCA) to calculate the principal components of the expression data using data from all genes. Euclidean distances between expression values are used. Samples are projected on the first two principal components (PCs) and the percent variance explained by those PCs are displayed along the x and y axes. Ideally your samples will cluster by group identifier.

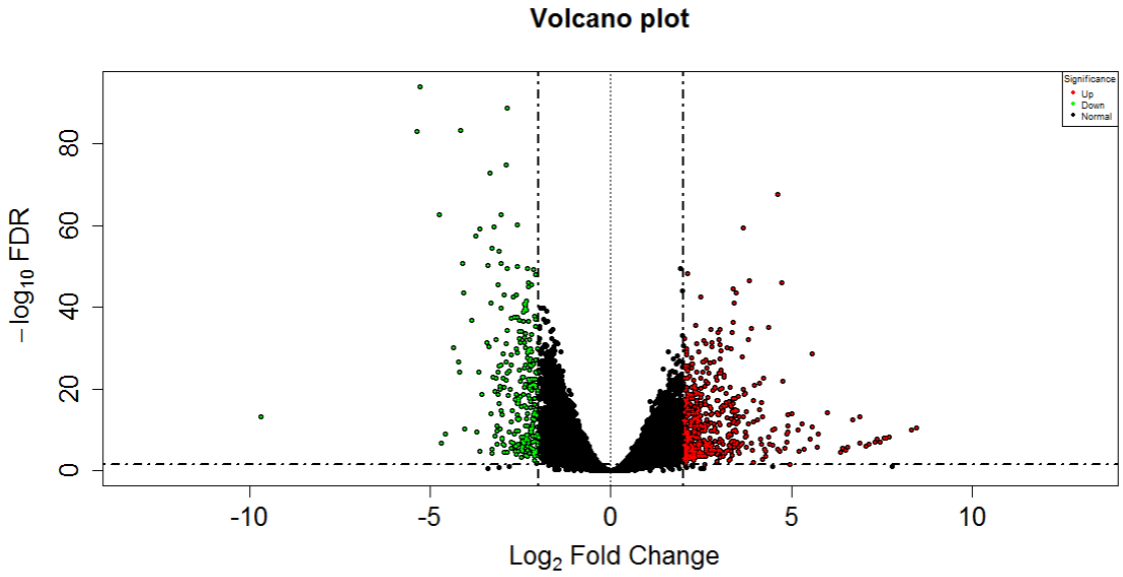


Analysis Plots

These plots use the p-values and fold changes to visualize your data.

Volcano Plot

This is a scatter plot log fold changes vs –log10(p-values) so that genes with the largest fold changes and smallest p-values are shown on the extreme top left and top right of the plot. Hover over points to see which gene is represented by each point.



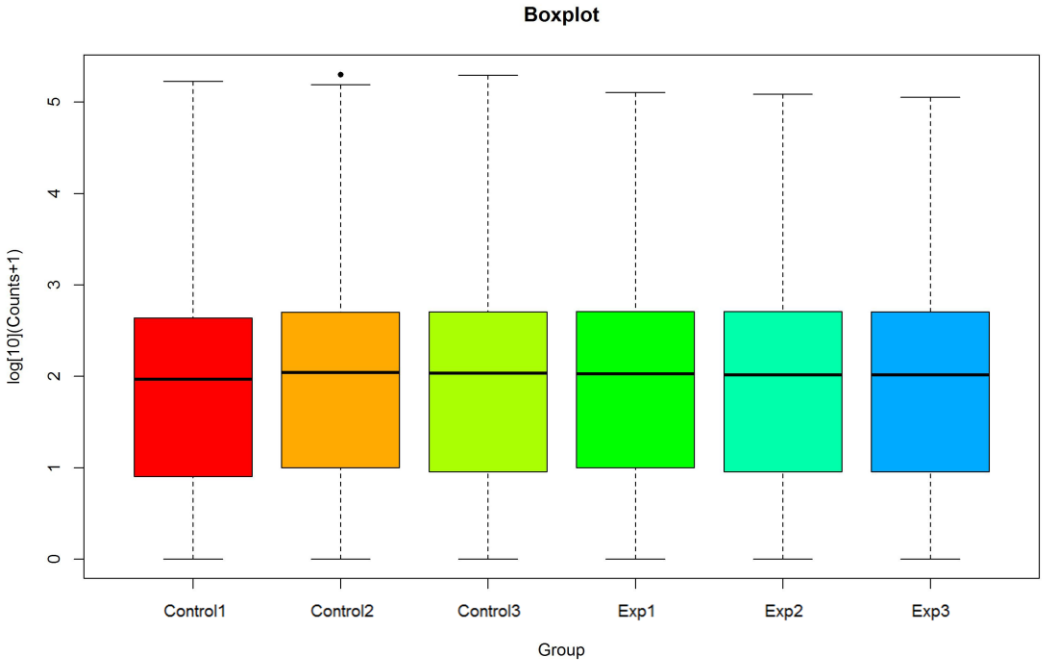
(https://en.wikipedia.org/wiki/Volcano_plot_(statistics)

Scatter Plot

This is a scatter plot of average gene expression in one group against another group. This allows the viewer to observe which genes have the largest differences between two groups. The smallest distances will be along the diagonal line, and points far away from the diagonal show the most differences. Hover over points to see which gene is represented by each point.

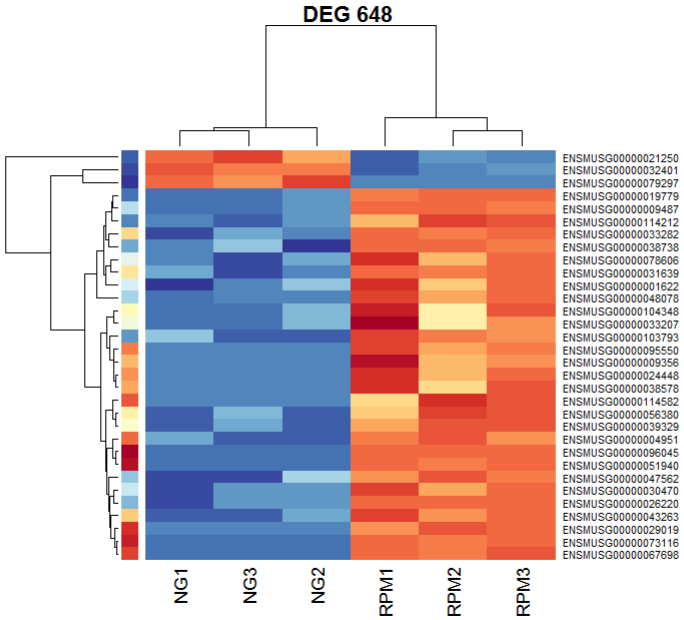
Gene Expression Boxplot

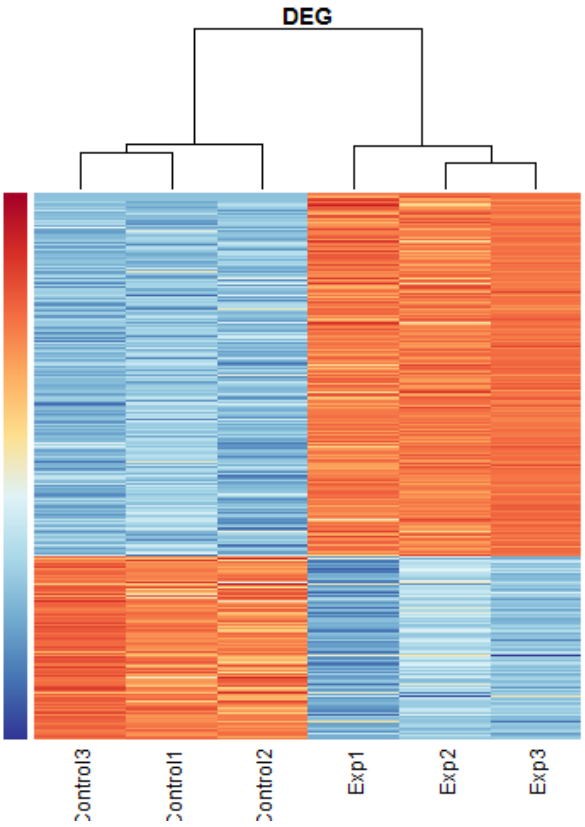
Use the search bar to look up genes in your data set. For selected gene(s) the stripchart (dotplot) and boxplots of the expression values are presented for each group. You may plot one or multiple genes along side each other.

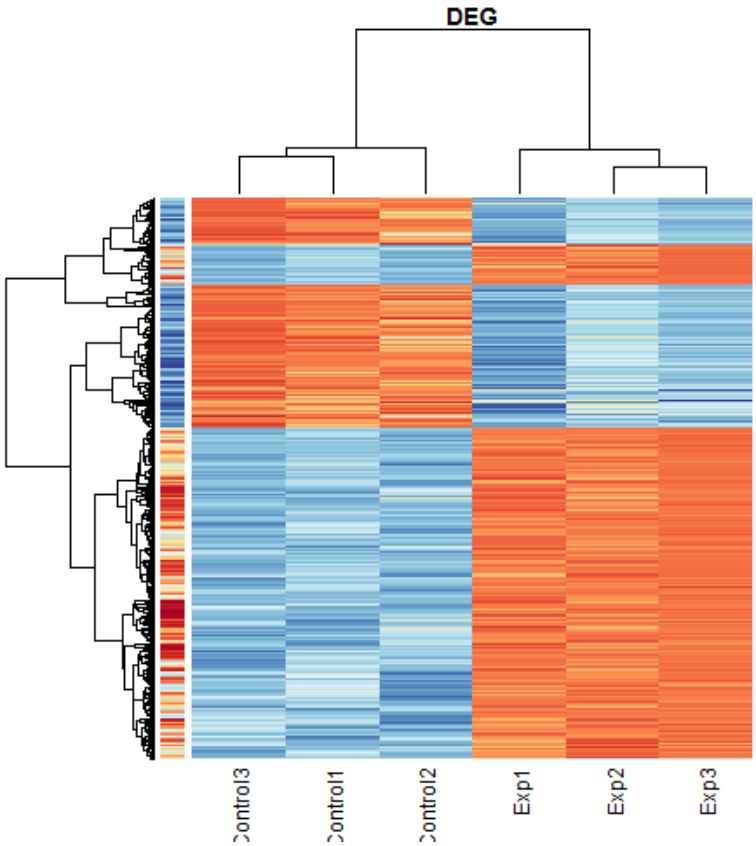


Heatmap

A heatmap of expression values are shown, with genes and samples arranged by unsupervised clustering. You may filter on test results as well as P-value cutoffs. By default the top 100 genes (with lowest P-values) are shown.




DEG
Control3
Control1
Control2
Exp1
Exp2
Exp3


DEG
Control3
Control1
Control2
Exp1
Exp2
Exp3