# Instructions for the START App

The app is hosted on shinyapps: https://kcvi.shinyapps.io/START/
Code can be found on github: https://github.com/jminnier/STARTapp

## Input Data

Either select the example data set or upload your own data file. After uploading a file and selecting the appropriate options, you must click the "Submit Data" button to populate the app's visualizations with your data.

### Example: RNA-Seq gene counts

This is a pre-loaded mouse RNA-seq example for exploring the app's features.

### RData from previous START upload

You may upload an .RData file that you previously downloaded from the START app.

### Upload Data

You may upload your data in two formats, raw counts, or analyzed data.

You must include at least one gene identifier or name.

Your file must have a header row.

The names of your expression/counts must be in the format: "group1_1" where "group1" is the name a group and "1" is the replicate id number. These must be separated by an underscore. The program determines the names of your groups from this column format.

#### Raw Data: Gene Counts

Raw counts contain read counts for each gene for each sample, along with gene identifiers.

***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 p-values are computed for each pairwise comparison. The "log2cpm" values are the expression values from the voom method.

Example file:
https://github.com/jminnier/STARTapp/blob/master/data/examplecounts_short.csv

**Analyzed Data**

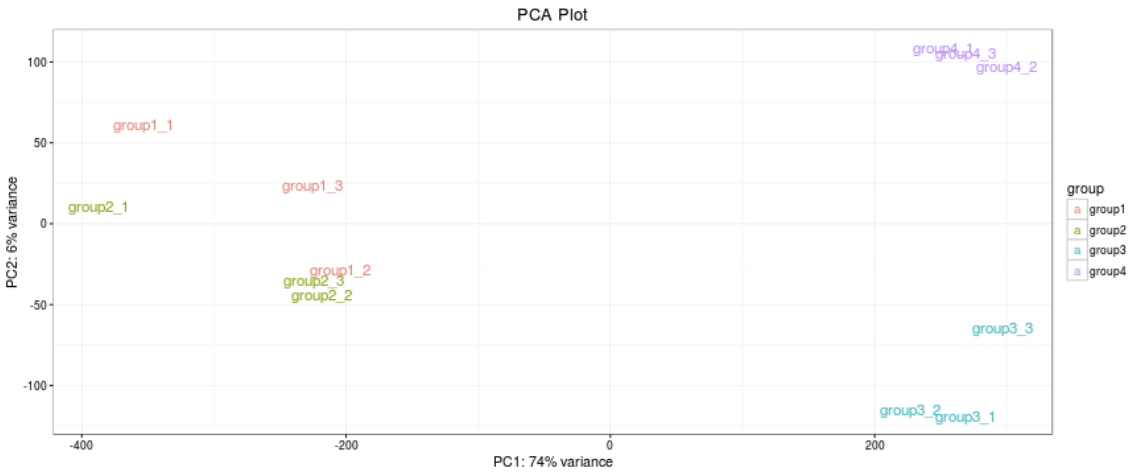
Analyzed data must contain some kind of expression measure for each sample (i.e. counts, normalized intensities, CPMs), and a set of p-values with corresponding fold changes for those p-values. For instance, if you have a p-value for the comparison of group1 vs group2, you can upload the observed fold change or log2(fold change) between group1 vs group2. 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.

Example file:
https://github.com/jminnier/STARTapp/blob/master/data/exampleanalysisres_short.csv

## Group Plots

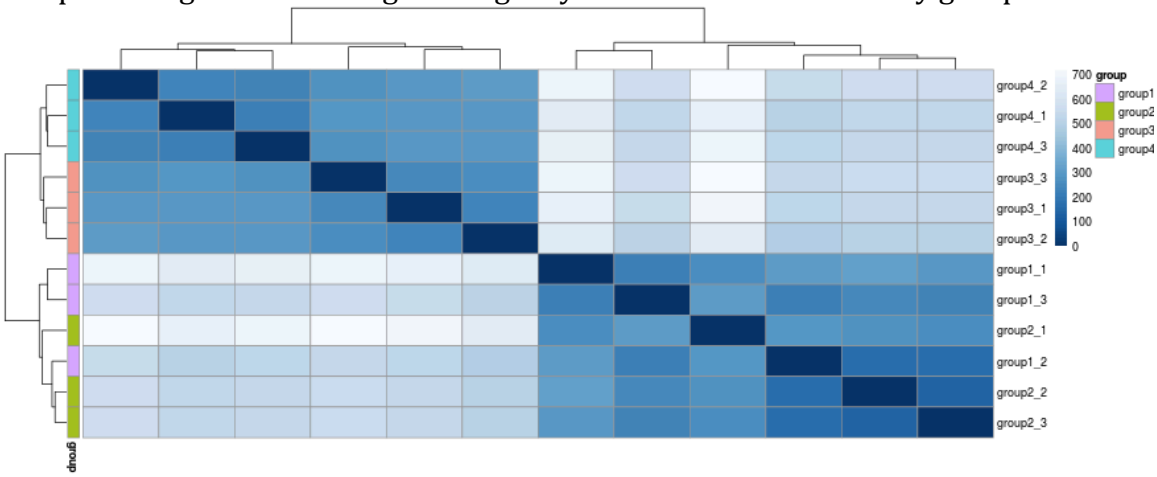
### PCA Plot

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.




## Sample Distance Heatmap

This plot displays unsupervised clustering of the Euclidean distances between samples using data from all genes. Again your data should cluster by group.

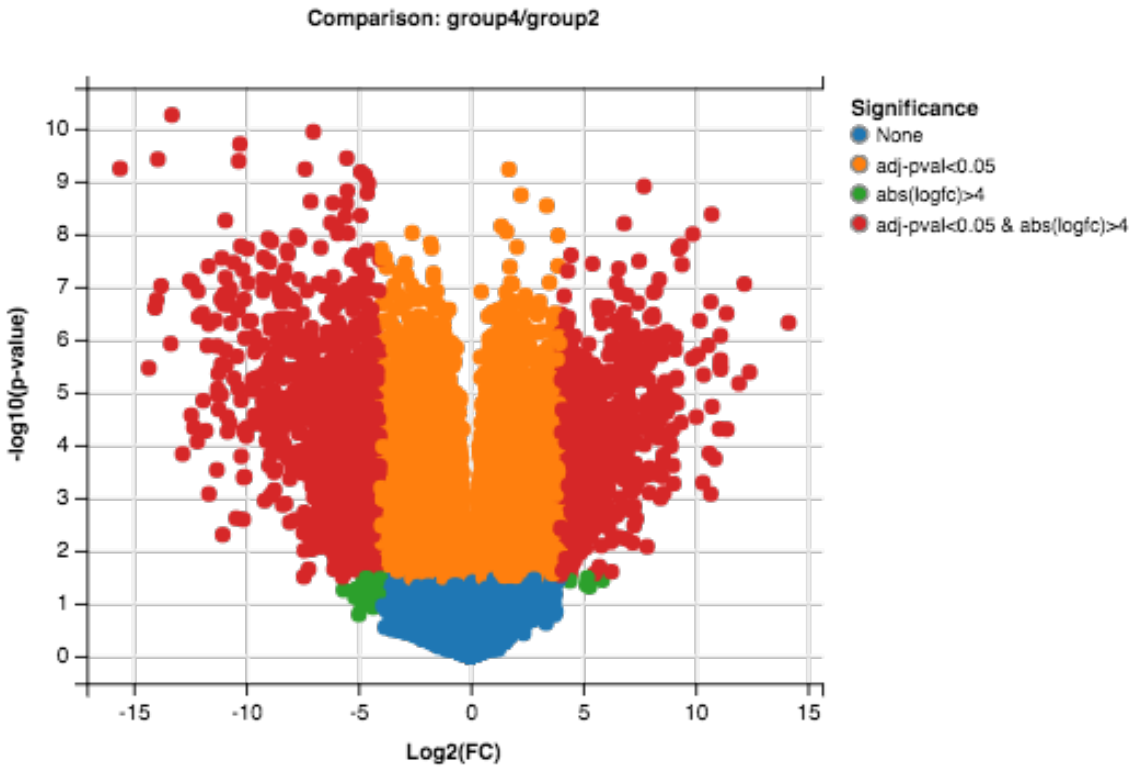



## Analysis Plots

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

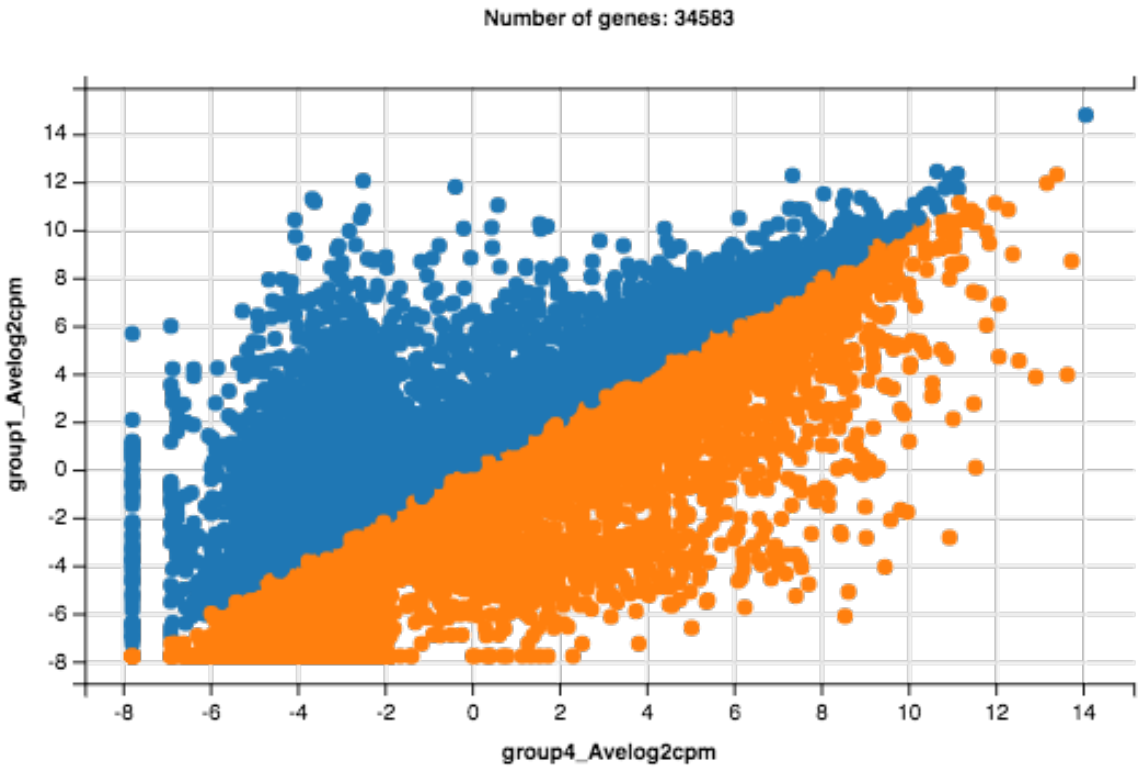
### Volcano Plot

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))



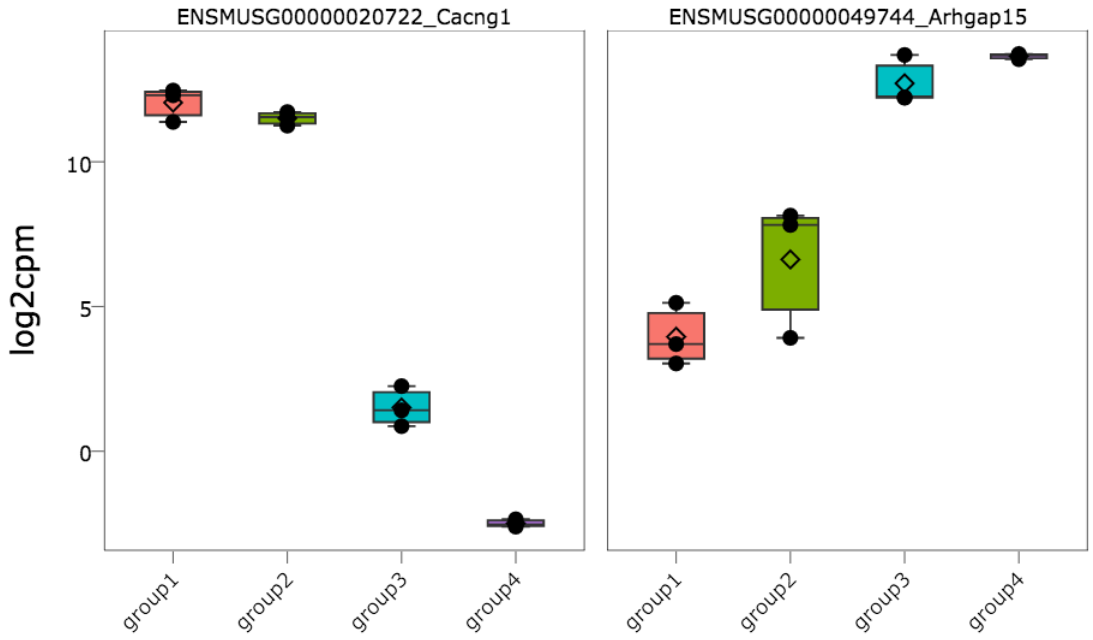

**Scatter Plot**

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.



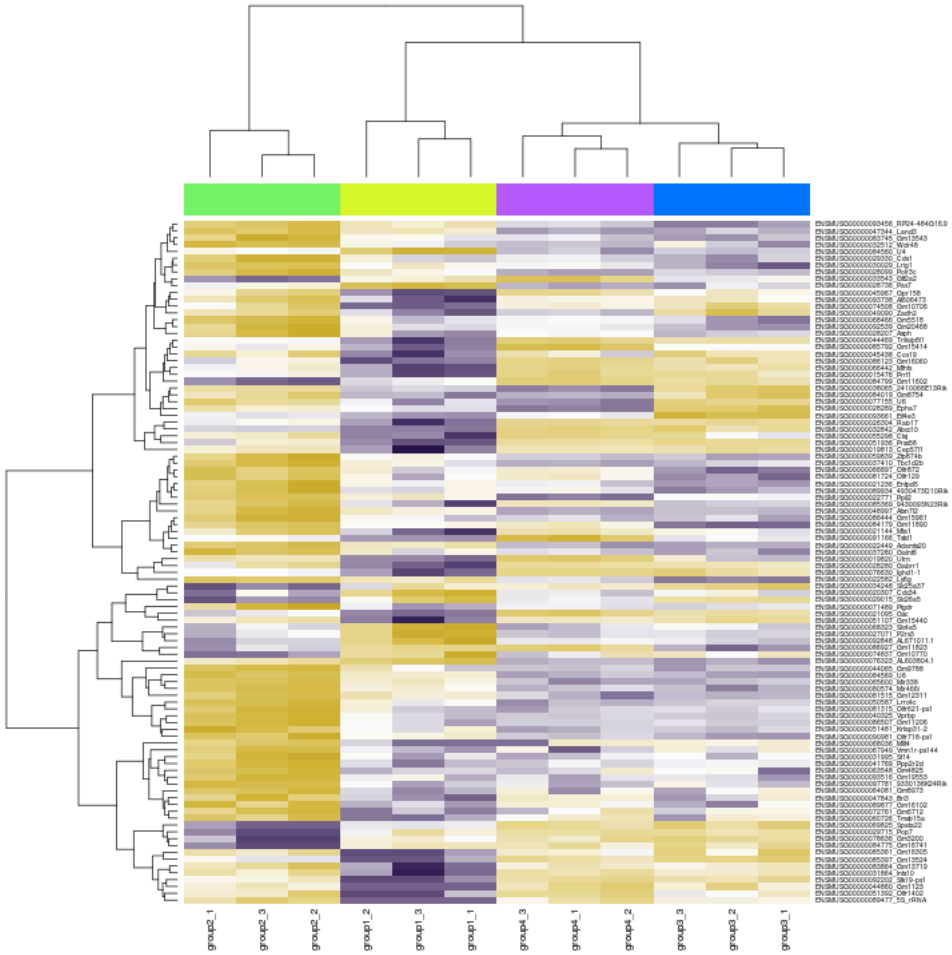

## Gene Expression Boxplot

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.


group2_1
group2_3
group2_2
group1_2
group1_3
group1_1
group4_3
group4_1
group4_2
group3_3
group3_2
group3_1