How to work with DESeq2

TERMINOLOGY
We use **samples** to denote units such as patients, persons, study participants, animals, plants, cells,...

We use **variables** to denote quantities that have been measured for each sample, such as gene expression levels, miRNA expression levels, protein concentrations, antibody concentrations, methylation levels, answers to questions in a questionnaire, ...

Normally a **data set** is described by a matrix where the columns represent samples and the rows represent variables.

INTRODUCTION
This how to document focus on analyzing count-based data with DESeq2 functionality. The implementation in Qlucore Omics Explorer is optimized for performance and is based on the publication "*Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.*", by Love, M. I., Huber, W., & Anders, S. (2014), Genome biology, 15, 1-21.

COUNT-BASED DATA
Several Instruments and techniques of data generation, such as RNA-seq, generate count-based data. To analyze such data there are multiple options:

1) Do a transformation (such as TMM) and use standard statistical methods such as ANOVA. This is supported in Qlucore Omics Explorer and has been for many generations.
2) Apply statistical tests specifically developed for count based data. One well known method in this field is DESeq2 - Differential gene expression analysis based on the negative binomial distribution.

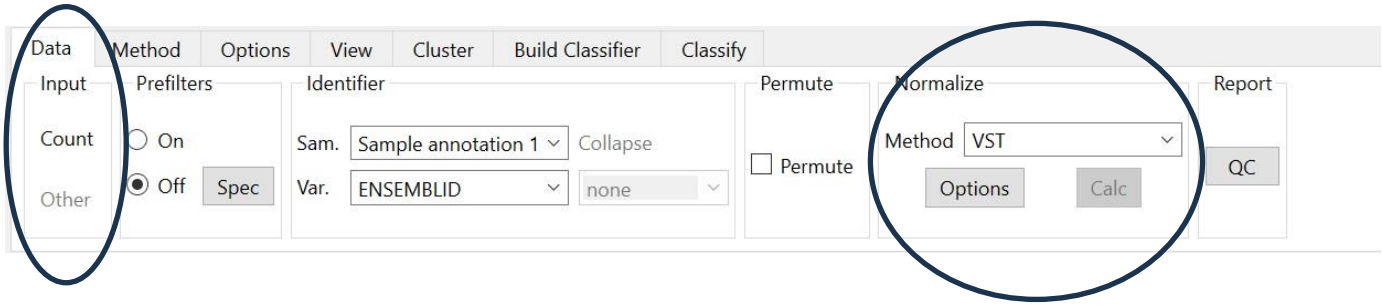
DESEQ2 FUNCTIONALITY
In Qlucore Omics Explorer (QOE) a broad suite of algorithms and tests are available for easy analysis of count data. The functionality includes:

1. Data import
2. VST normalization
3. A suite of statistical tests for count data
4. Independent filtering

All other functionality in the program is available too.

DATA IMPORT

The data import is done either directly from the open bam files dialog (File menu) or by using the Wizard (File menu). In the Data tab it is indicated what type of data set that is imported, and you also choose the type of Normalization.



VARIANCE STABILIZATION TRANSFORMATION (VST)

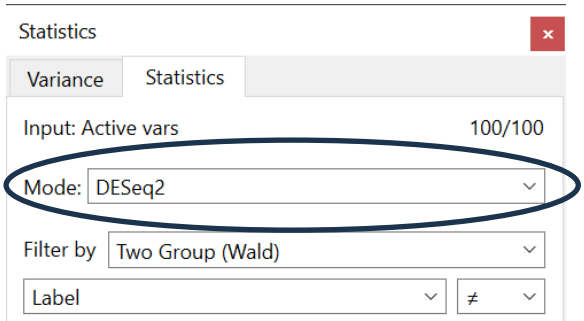
VST normalization is used to enable visualization of data in for instance PCA and other plots. Note that other Normalization options are available, but we recommend VST for count-data.

STATISTICAL TESTS

The below listed statistical tests are dedicated to count-based data.

- **Wald**
  - two group
  - linear regression
  - rank regression
  - paired two group
- **Likelihood ratio test (LRT)**
  - two group
  - multi group
  - linear regression
  - quadratic regression
  - rank regression
  - paired two group

If you are working with DESeq2 based statistics, you make that choice at the top part of the statistics dialog.



Note: That the Variance filtering is independent and applied prior to the statistical tests, regardless of it is Standard statistics or DESeq2 tests.

INDEPENDENT FILTERING
Independent filtering enables removal of variables to improve the q-value.

HOW TO FIND VARIABLES THAT DISCRIMINATE TWO OR MORE GROUPS

Now, let us work through an example: "How to find variables that discriminate two or more groups."

This is precisely what multi group comparison is designed to do. In QOE there are several different tests that can be applied to find discriminating variables, depending on the experimental design (DESeq2 (LRT), t-test, Welch, ANOVA, Kruskal-Wallis,…). In this example (which is based on the example count data file from the Help menu – "COVID RNAseq Count Data Set", we will use the Multi Group Comparison and the LRT test. You can read about the statistical concepts in the reference manual.

Note: If you have two groups, use a two-group comparison test.

EXAMPLE WORKFLOW

WE WILL WORK THROUGH THE FOLLOWING STEPS

1. Load data
2. Do normalization
3. Select statistics
4. Visualize data in a PCA plot
5. Inspect variables in a synchronized variable PCA plot
6. View data in a heatmap
7. Export a list of discriminating variables
8. Make a selection of a subset of variables
9. Generate box plots
10. Discuss fold change

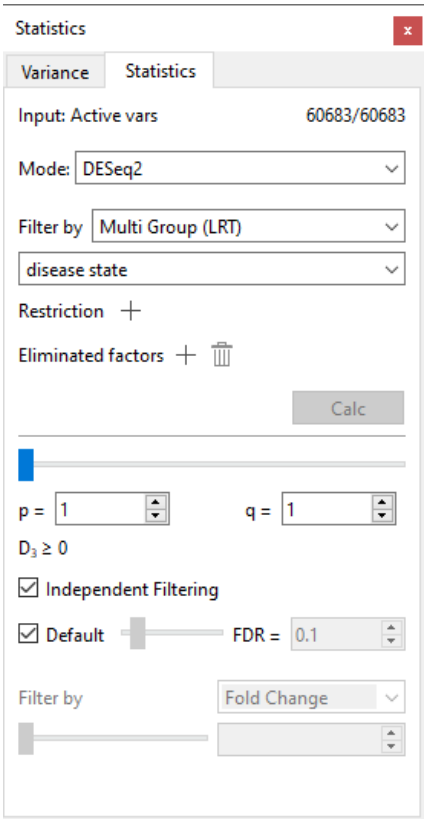
STEPS

Load you data if you have not already done it (File menu). Also apply VST Normalization (Data tab) if it is not already done, and press Calc.

Note: The COVID example data set is found in the Help menu.

If you are working with your own data set, make sure that you have the subgroups you would like to discriminate, described by a sample annotation. If the clinical information or group information required is not in the program yet, see for instance the "How to add a sample annotation" document for more information. In the example we will use the sample annotation "disease state".

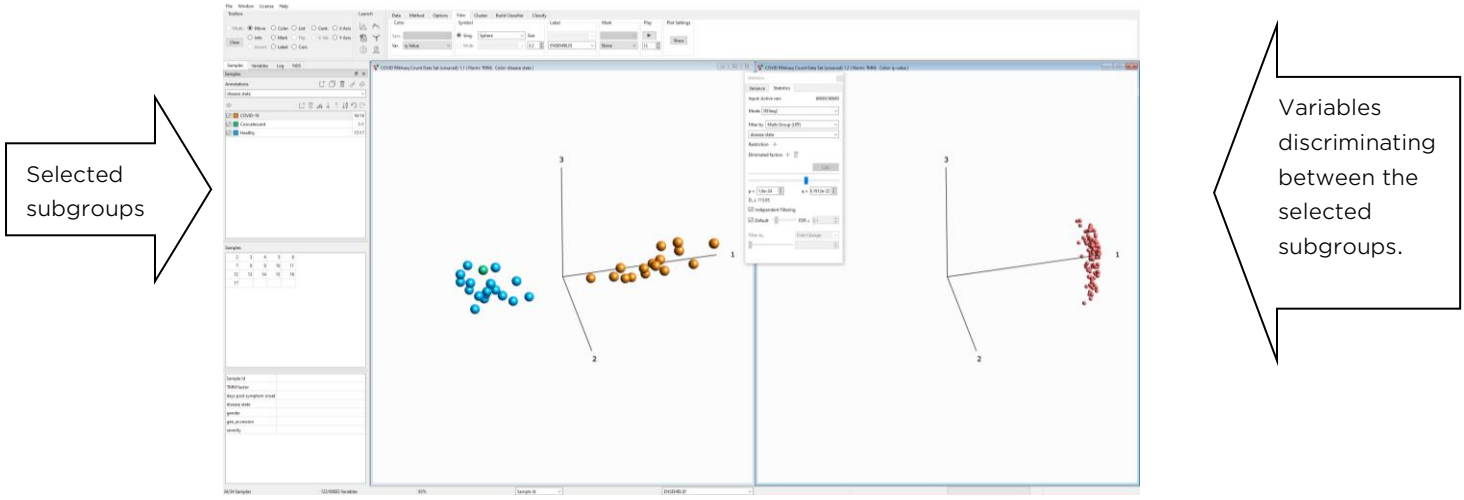
Select DESeq2 statistics at the top of the statistics dialog. Select Filter using Multi Group (LRT) and the subgroup that you are working with (in our example "disease state"). Press Calc.



When the calculation is done the sliders are available to focus on the most discriminating variables. In the figure, we have two synchronized plots showing the selected sample groups to the left and the variables to the right[1]. Through this plot set-up you get a good overview. In the example we use PCA plots but you could also have used for instance a heatmap.

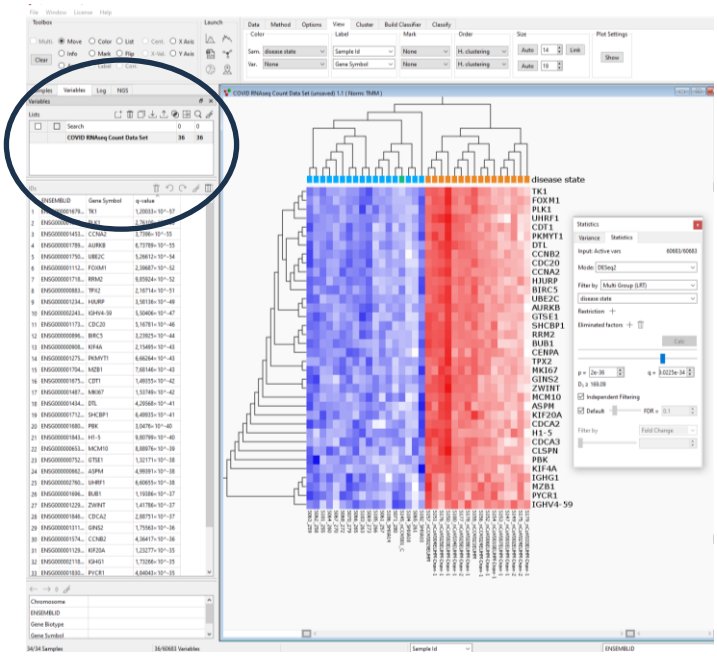
[1] To open a synchronized window you need to select the Window menu and then select synchronized plot. Finally select the Window menu again and select menu item tile (Ctrl – T) to show both windows.

The variables in the plot are in this example colored after their q-values.



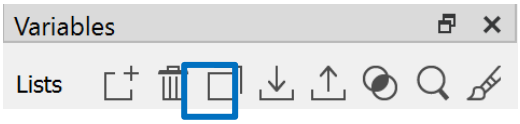
Note that the green sample (labelled as “Convalescent”) in the PCA plot to the left is as expected grouping with the blue samples (labelled as “Healthy”).

The selected variables are available in the automatic list in the variable tab. The filter settings



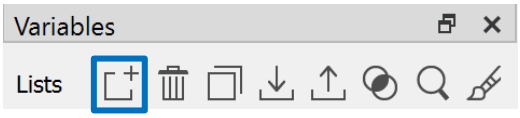
keep in this example 36 variables, see the encircled area. This list is easily exported.

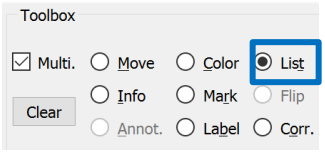
Select the List export icon.



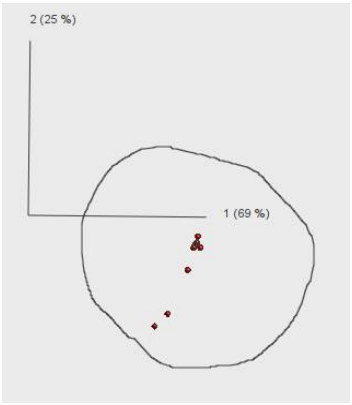
LIST TOOL

If you are interested only of a few variables it is easy to use the list tool (Toolbox) it is easy to select the variables of highest interest directly in a plot and create a list. Select the New button to first create the list and then add variables to the list with the list tool.





Draw a closed curve clockwise around some of the variable that are of interest. You do this by holding down the left mouse button while at the same time moving the mouse tool clockwise around the selected genes to create a closed curve.



Note: Selections can also be done in the heatmap by clicking in the dendrogram.

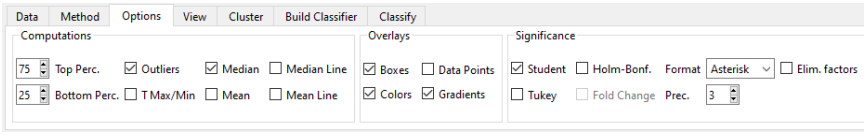
The created list is now available in the Variable list panel. The list can be populated with for instance p-value and q-value.

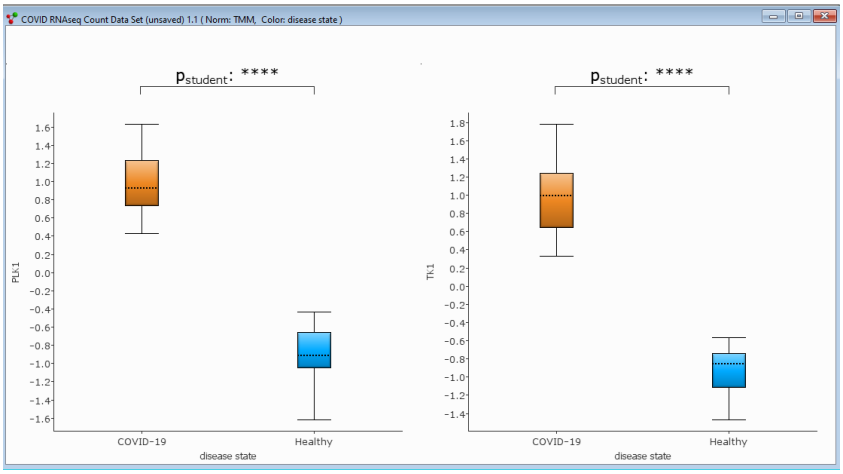
BOX (OR VIOLIN) PLOT

When you have created a limited selection of variables you can use scatter violin or box plots to visualize the results further. The below box (violin) plot is created by the following steps

- Uncheck the Convalescent group in the Sample tab since there is only one sample in this group. Press calc again.
- Close the Variable PCA plots
- Select the plot type “Box”(“Violin”) for the remaining Sample PCA plot
- Select the Leukemia Subtype annotation as the annotation for the X-axis.
- Select variables from created variable list with the Y-Axis tool.

In the plot the two selected variables (genes) are displayed. They were selected among the most discriminating genes In the Options the tab there are a rich set of options to configure the plot.





FOLD CHANGE

When using two group comparison to determine variables that discriminates groups the Fold Change[2] filtering option is also available. Fold Change is a metric for the effect.

Note: In DESeq2 (R), there is a parameter called `lfcThreshold`, which does not correspond to filtering by log fold change. Instead, it adjusts the null hypothesis (typically from $\beta \neq 0$ to $|\beta| > \text{lfcThreshold}$, but it depends on test direction) and it does not actually filter the variables.

DISCLAIMERS

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.

[2] For two-group statistical tests, the Fold Change for a variable is calculated from the difference between the arithmetic average over the first group and the average over the second group. The difference δ is interpreted as the 2-logarithm of the Fold Change, and therefore, the Fold Change is calculated as 2^{δ}. Please note that the data must be log transformed for a correct calculation of the Fold change! This can be performed in the Data tab.