How to work with a dataset with few samples

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. By a data set with few samples we mean a data set with 4 to 20 samples.

HOW TO WORK WITH A DATASET WITH FEW SAMPLES

When working with a dataset with a limited number of samples (but still more than two) the relevant analyses and investigations are often carried out in the variable space. We will in this document briefly describe some useful analysis steps. The specific analysis will be dependent on the specific experiment.

We will be using a data set with 12 samples. The samples are divided into 4 groups of triplets, corresponding to biological replicates. One group is the control group and the other three groups have been stimulated in different ways.

The initial plot of data looks is seen in Figure 1.

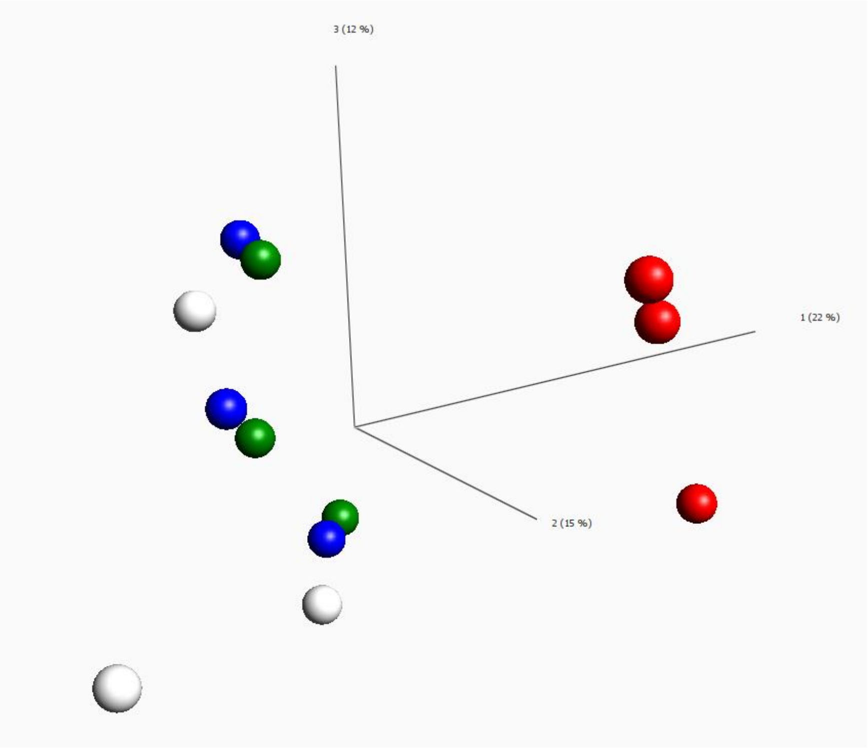


Figure 1: *A first inspection of the samples*

The red group is the control group. We can clearly see that the control group is different from the three stimulated groups, while three stimulated groups are more similar to each other.

Coloring the data according to the available annotations we observe that the annotation patient ID is important to consider since there are big individual differences between patients.

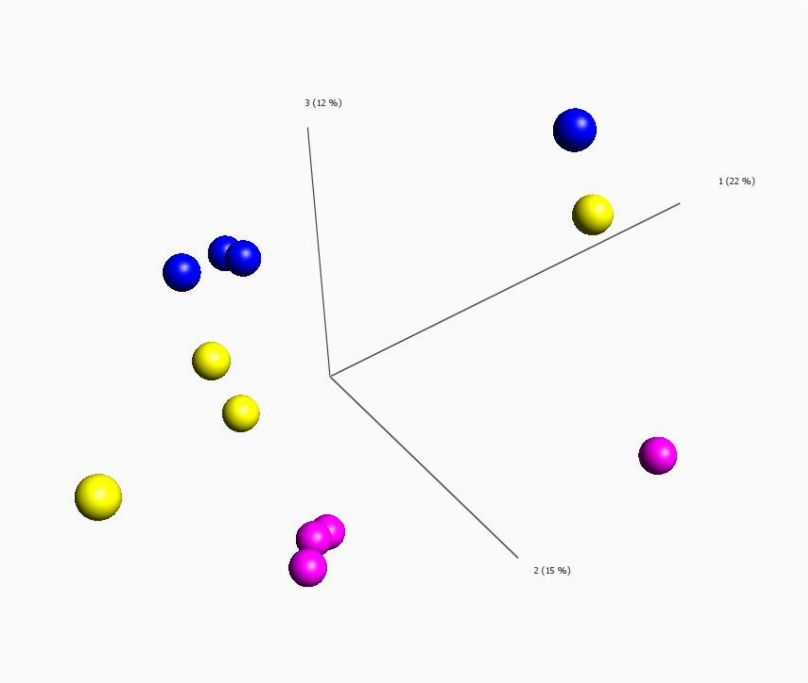
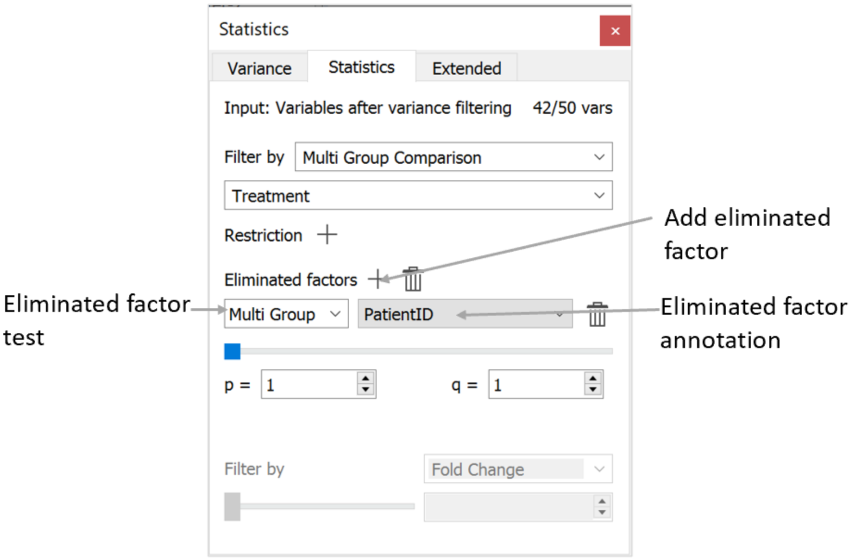


Figure 2: *Colored according to the individual patients*

The first step in this analysis will be to take into account that there are differences between the individual patients and remove this effect before we analyze the different stimulations.
This can be done by using **the Eliminated factors** function in the statistics dialog. The factor we eliminate is in this example called patient.



Statistics

Variance | Statistics | Extended

Input: Variables after variance filtering 42/50 vars

Filter by: Multi Group Comparison

Treatment

Restriction

Eliminated factors

Multi Group | PatientID

p = 1   q = 1

Filter by: Fold Change

We can now, in Figure 3, see that the three stimulated groups (green, blue and white) are clearer separated and that the samples in each group are closer to each other.

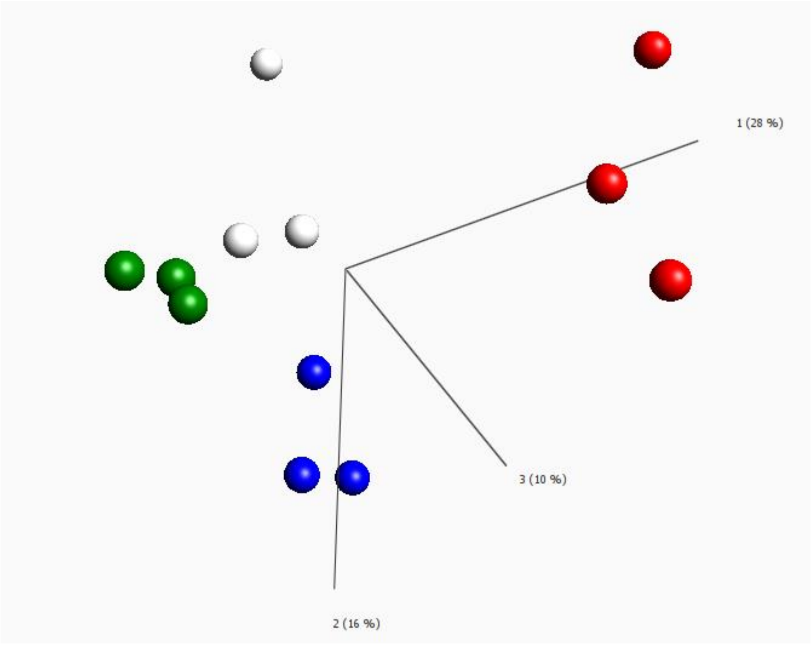


Figure 3: *Samples after eliminating the patient effect*

Since the groups are known in this dataset the questions that we want to work with are primarily related to the variables.

The variables are presented in a Variable PCA plot, shown in Figure 4. We have reduced the set of active variables to those 300 variables that are most highly discriminating between the four stimulus groups, using the p-value slider (based on a Multi Group Comparison (ANOVA)). In the plot the variables are colored according to their values in the green sample group.

Later in this document, we will show how you can find variables that correlate strongly with one of the remaining variables. You could also use other types of biologically guided analyses to find out more about the remaining variables.

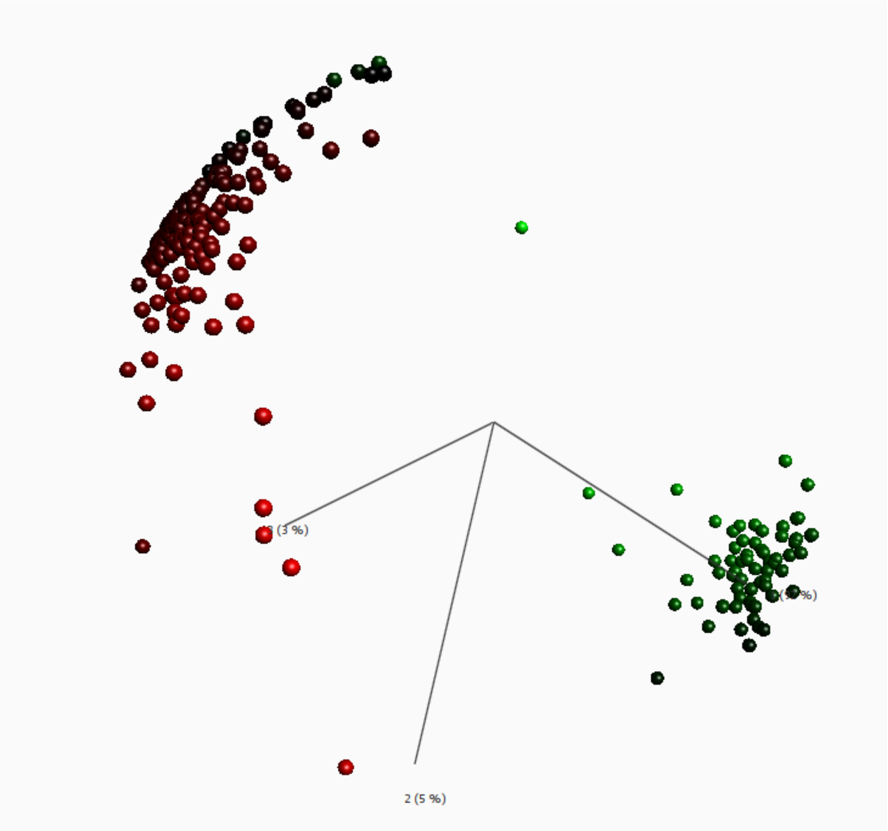


Figure 4: *Variable plot. First inspection.*

Next, we will investigate which variables that discriminate the three stimulated groups from the control group. Now, we compare two groups of samples to each other and therefore, instead of using a Multi Group Comparison we will use a Two Group Comparison (t-test). An alternative would be to use a Welch test. We select Two Group Comparison and the annotation Treatment, with the level Normal Control, in the statistics dialog. Filtering until around 100 variables remain gives the plot shown in Figure 5. In this step we use a heatmap to visualize the result since it give a good overview. You could have used a Variable PCA plot too. Note that all 100 variables are not visible.

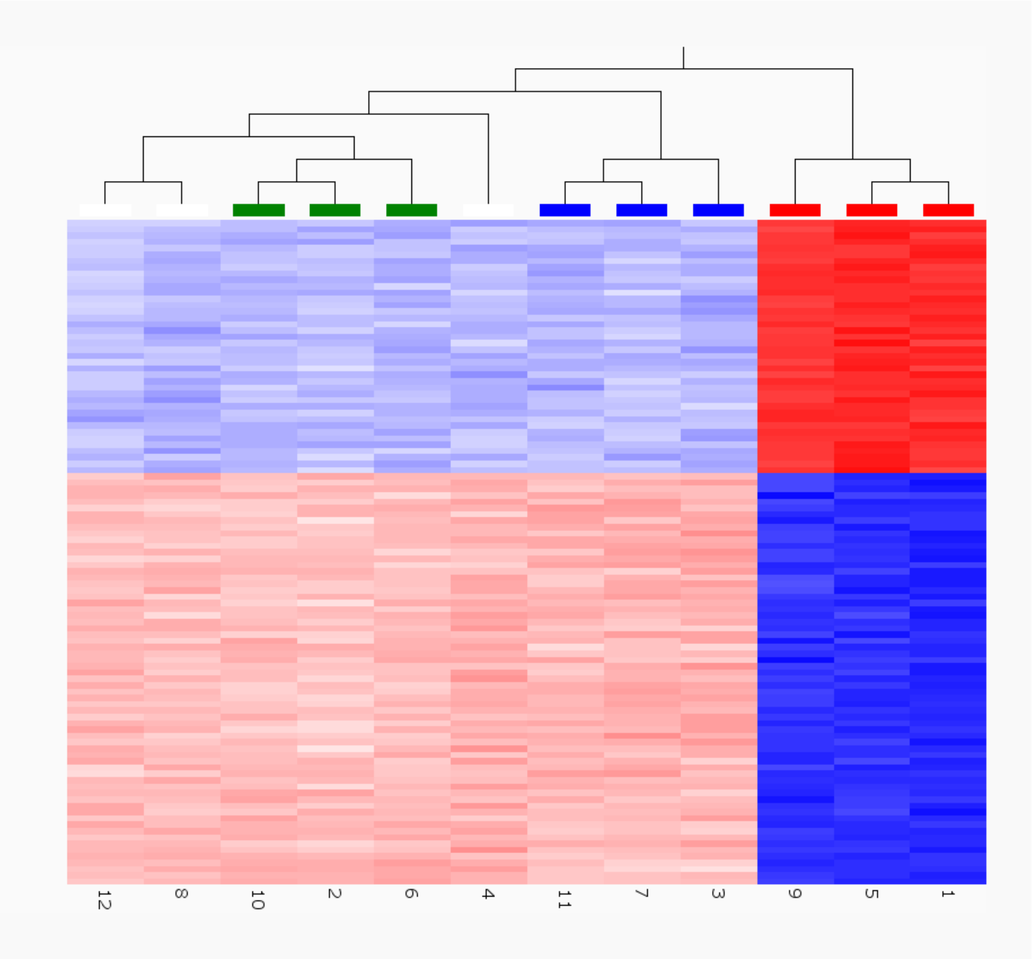
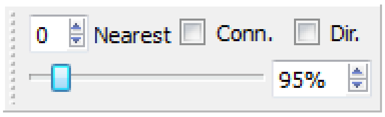


Figure 5: *Variables separating the control group from the stimulated groups.*

The next thing we will do is to find those variables that discriminate between the three stimulated groups. Deselect the control group by using the active samples button. Select Multi Group Comparison in the statistics dialog and filter by using the p-value slider.

To get a better overview it is good to use synchronized plots. The locations are synchronized so that the variables that are up-regulated for a group of samples are located in the same point in space as the group of samples. The variables in the variable plot are colored by their values in the green sample group. We have further connected all variables in the plot that have a correlation of at least 95%, see the network toolbox.



0 Nearest □ Conn. □ Dir.

95%

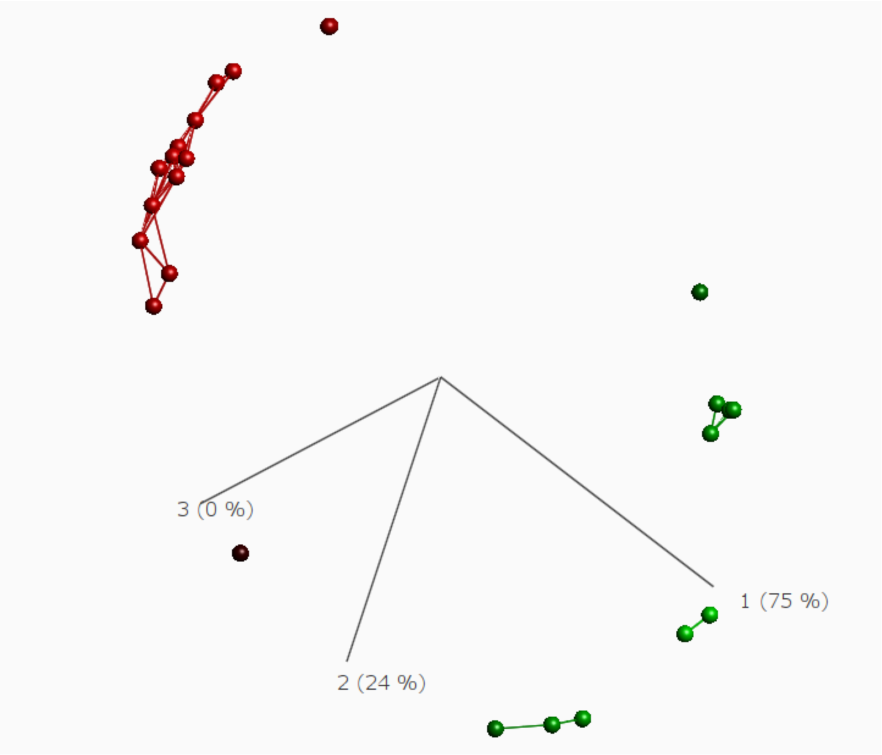


Figure 6: *Variable networks created by joining pairs of active variables that are highly correlated.*

The last step of this example is to look in detail on all the highly correlated variables shown in Figure 6. We use multiple scatter plots to demonstrate this. To generate the plot, we performed the following steps

- Create a variable list of the highly correlated genes. The list includes 12 genes.
- Use a sample scatter plot
- Populate the x-axis according to the annotation Group
- Populate the y-axis by using the Y-Axis tool and select the created list

The result is presented in Figure 6.

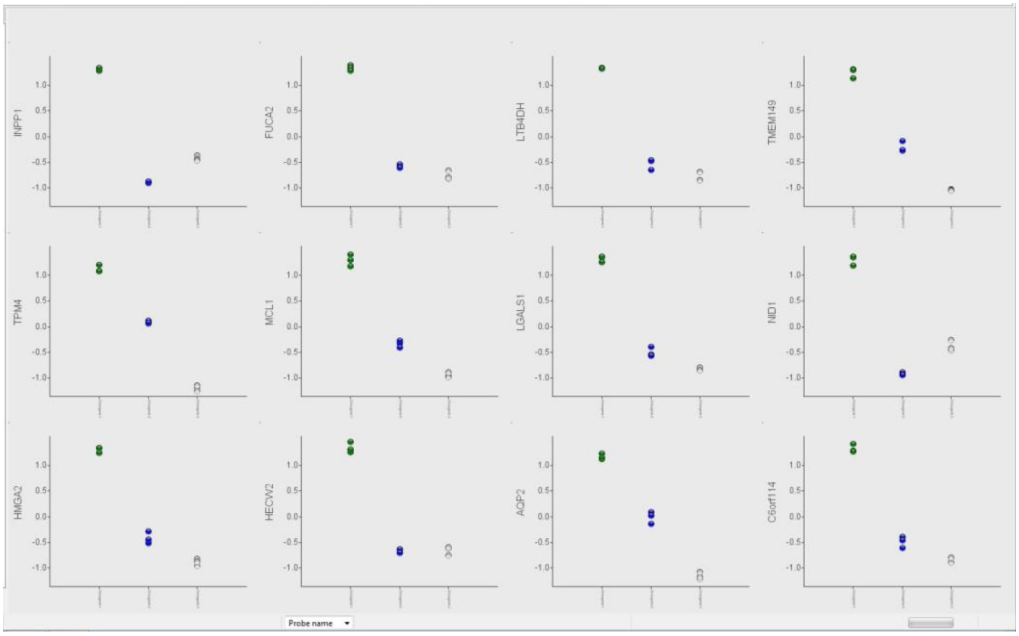


Figure 6: *Scatter plots of 12 variables*

DISCLAIMER
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.