How to work with Flow Cytometry data

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 document outlines a few different ways to use Qlucore Omics Explorer (QOE) for flow cytometry data analysis. We will show how to use visualization and statistics tools in QOE to find outliers and subgroups among samples, along with discriminating features.

As a starting point for the analyses presented here, we need a set of features (variables) describing each of the samples.

A comprehensive description of all functions can be found in the Reference Manual that is supplied in the Help Menu of Qlucore Omics Explorer. On www.qlucore.com you can find more documentation and watch instruction videos.

EXPORTING SAMPLE FEATURES TO QLUCORE OMICS EXPLORER

The easiest way to export features (variables) describing each sample from gating software to QOE is by using .csv files. Counts of different populations is often a natural choice of features to export. However, any feature varying in a relevant way between samples can be used, for examples relative frequencies, mean fluorescence intensities (MFI's) or MFI ratios.

You need to ensure that either each column or each row corresponds to a sample. Sample annotations can be loaded separately into Qlucore Omics Explorer (File > Import > Sample Annotations). Make sure that the sample names in the feature .csv file are in accordance with the sample names in the annotation file.

If you use FlowJo (http://www.flowjo.com) for manual gating you can use a synchronized group to do batch analysis of all samples (http://docs.flowjo.com/d2/workspaces-and-samples/ws-groups/ws-gatecopying/). Then use the Table Editor (http://docs.flowjo.com/d2/tabular-reports/) to define interesting features for export to a .csv file. To set the gates, you can either use one typical sample as a template, or you can first

concatenate all the sample files into a new .fcs file and do the gating on the concatenated file. Using a concatenated file allows you to consider the data in all samples simultaneously.

LOAD DATA TO QLUCORE OMICS EXPLORER

In this document we use a data set with leukemic and normal samples, stained with five colors (Aghaeepour et al. 2013). The data is publicly available from FlowRepository (Spidlen J et al. 2012). Each sample has been analyzed with eight tubes, including isotope control. Here we look at the first 100 samples from tube 6 (i.e. 0006.FCS, 0014.FCS, 0022.FCS, …).

The data can be downloaded from https://flowrepository.org/id/FR-FCM-ZZYA. A script (AMLdownload.R) doing this automatically using R Bioconductor is available at the Qlucore support pages – search for Flow cytometry or ID 111.

Counts of lymphocytes, monocytes and neutrophils, as well as some lymphocyte subpopulations have been extracted by manual gating and exported to a .csv-file.

We first load the count data into Qlucore Omics Explorer with **File > Open With Wizard**. Then we load sample annotations from the file *AML sample annotations n100 LMD ID.csv* (available at the support pages (ID 111) on www.qlucore.com) using **File > Import > Sample Annotations**.

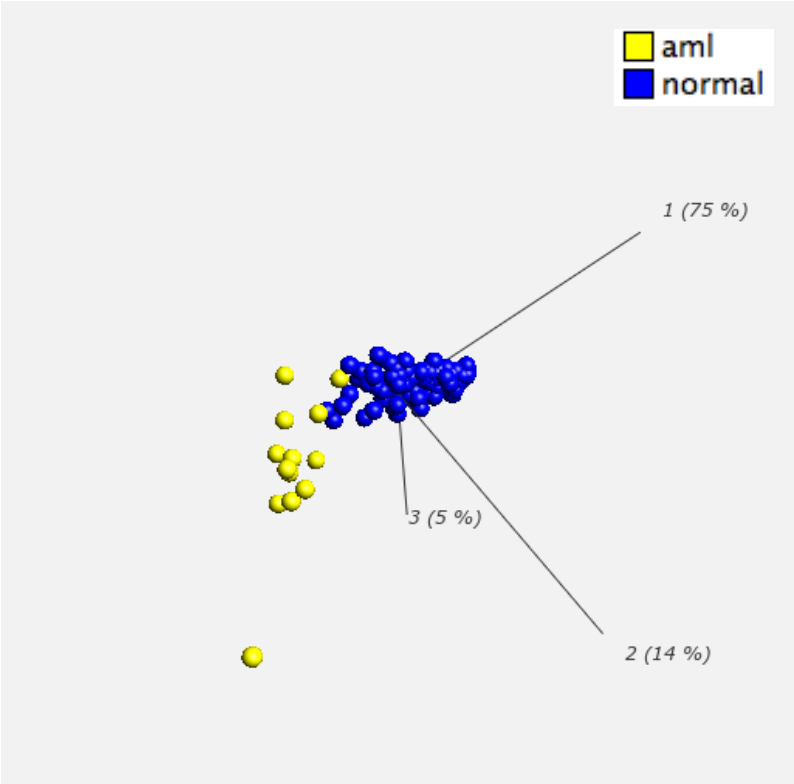
NORMALIZATION

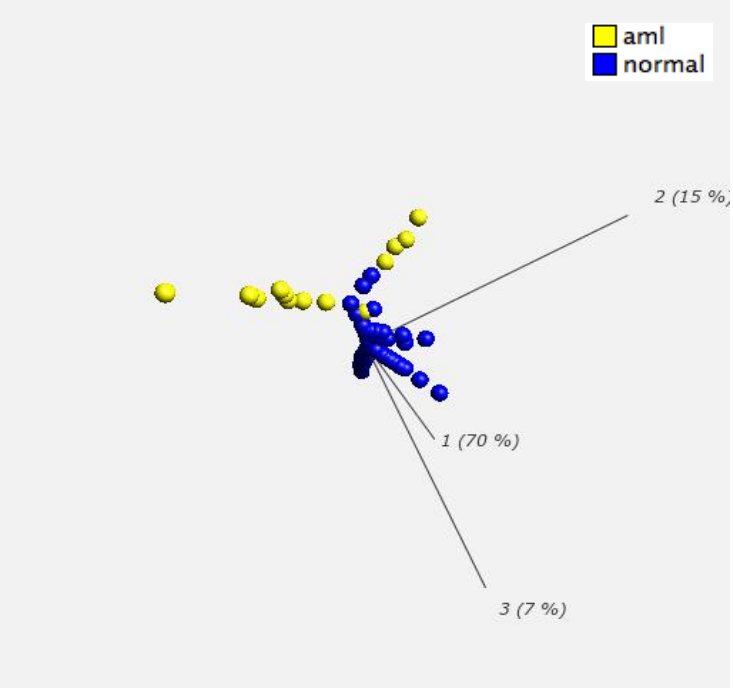
After loading the data, you need to decide which normalization to use and if you should log-transform your data. These decisions should be based on which features you use and what type of patterns/deviations you want to detect in your data. General guidelines are given in the Appendix. Since our example is based on count data we do not scale the variables to unit variance. This means that we go to the **Method** tab and choose **Mean=0** under Normalization.Now we are ready to explore the data in Qlucore Omics Explorer.

PRINCIPAL COMPONENT ANALYSIS (PCA)

In the **Method** tab choose **PCA** under **Plot Type**. In the Samples dock window, select the ”Condition” annotation and click on **Color samples**.

We can see that the normal samples are clustered tightly together, whereas the AML samples have much more variation. We can also immediately see one potential outlier among the AML samples.





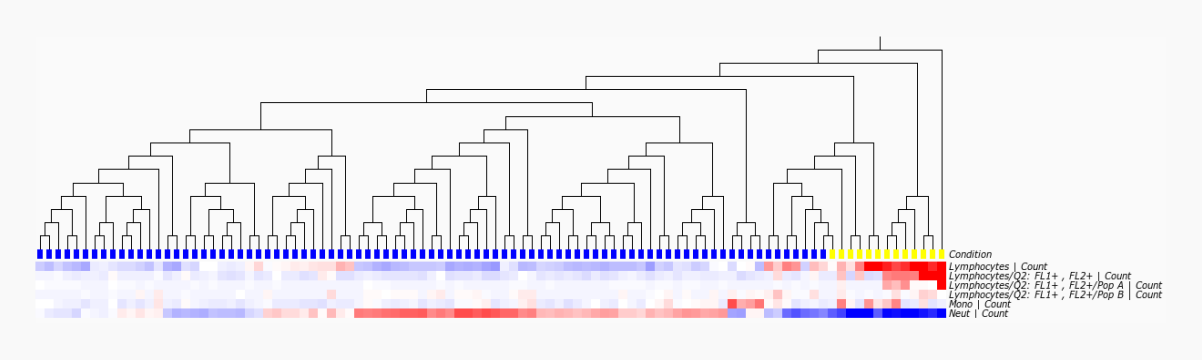
ISOMAP

Go to the **Options** tab. Under Isomap, click on **Set**. Isomap uses local distances between samples in the high-dimensional feature space to create a three-dimensional visualization. This enables us to see that there are two clear groups among the AML samples.

HEAT MAP AND HIERARCHICAL CLUSTERING

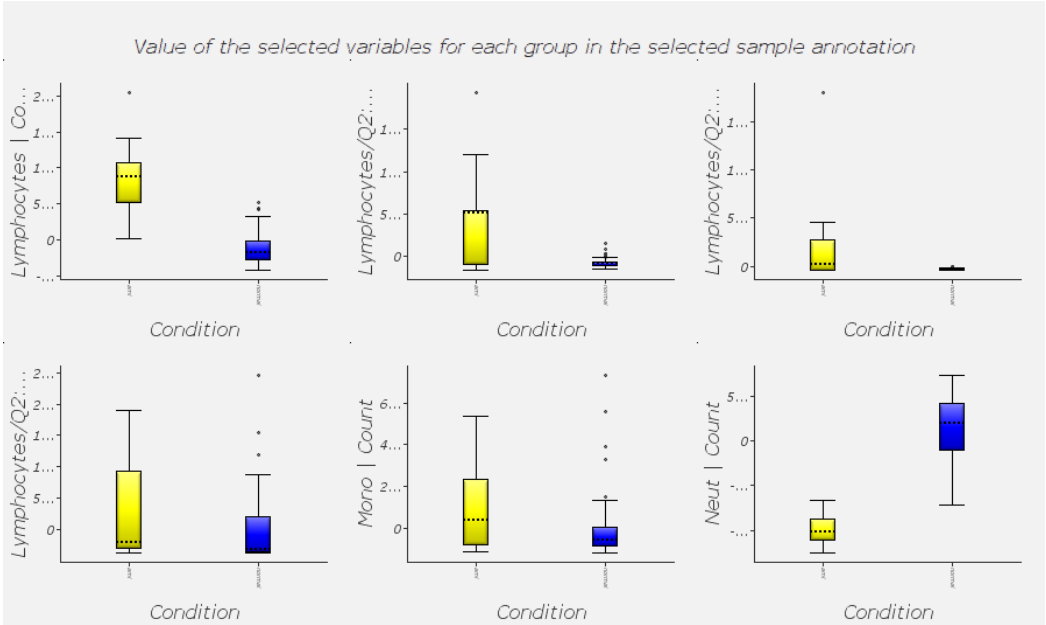
We can also visualize the difference between normal and AML samples in a heat map with hierarchical clustering. In the **Method** tab select **Heat** under **Plot Type**. Go to the View tab. Under **Order** set Sample **Order** to **Hierarchical Clustering**. To see which samples are annotated as AML samples, in the Color box set **Sample Color** to **By the annotation** — ”Condition”. The hierarchical clustering is visualized with a dendrogram shown above the heat map. Note that in each split in the dendrogram the ordering, i.e. which cluster that is to the left and which cluster that is to the right, is arbitrary. You can swap this by selecting **Flip** under **Toolbox** and click on the junctions in the dendrogram.

For the AML data set we obtain the heat map below. The AML samples (yellow) are to the right. We can see that in general, the AML samples have much higher lymphocyte counts and much lower neutrophil counts than the average.



BOX PLOTS

To further analyze how the counts vary between groups we can do box plots. In the **Method** tab, select **Box** under **Plot Type**. Under **Axis Data Selection** select ”Condition” for the X axis. Then choose “Variables selected by the Y axis tool” for the Y axis and click on the file name in the **Variable** dock window. We get box plots of each variable (feature) for each condition, shown below. To find the variables which have a statistically significant difference between the two groups we can use the **Statistics** dock window as described below.



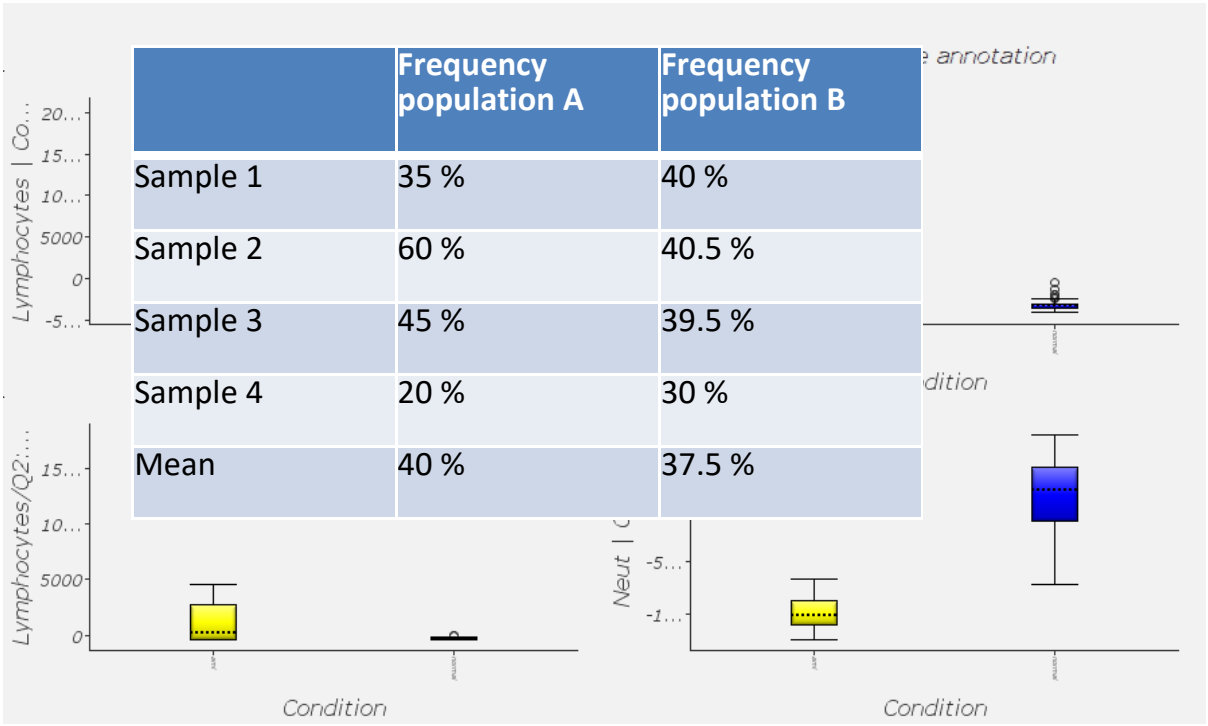
STATISTICAL TESTS

To only show variables that show statistically significant variation between the two groups, we can use the **Statistics** dock window to do a t-test. This can be combined with any of the plots described above.

Open the **Statistics** dock window through **View > Dock Windows > Statistics**. Select **Filter by Two Group Comparison**. To specify the annotation that distinguishes groups, for the AML data set we select ”Condition”. We choose significance level 0.05 by setting **p = 0.05**, meaning that all variables with a p-value below 0.05 are displayed. At this significance level all variables have a statistically significant difference between groups. At **p = 0.001** we have four significant variables, this is shown below for the box plots. We can also drag the slide to continuously change the p-value bound.

Note: The p-value is not corrected for multiple testing. If you have a large number of features it makes sense to use the q-value instead since it compensates for multiple tests based on the False Discovery rate.

A comprehensive description of the statistical tests available in Qlucore Omics Explorer can be found in the **Reference Manual** in the **Help menu.**



	Frequency population A	Frequency population B
Sample 1	35 %	40 %
Sample 2	60 %	40.5 %
Sample 3	45 %	39.5 %
Sample 4	20 %	30 %
Mean	40 %	37.5 %

APPENDIX: HOW TO CHOOSE NORMALIZATION AND LOG-TRANSFORMATION

NORMALIZATION

Normalization is chosen in the **Methods** tab IN QOE. It impacts for example PCA plots, heat maps, clustering and the y-axis of box plots. For flow cytometry data you should always use Mean = 0, and in many cases you should also use Var = 1.

The decision about which normalization to use should be based on the types of features that you have extracted from your data and how you value variations in data. The following three points need to be taken into consideration when deciding whether to use Var = 1:

- If you have features whose values are not comparable with each other, for example if you have both MFI's and counts as features, you need to use Var = 1.

- Using Var = 1 means that you will consider how extreme each feature is in each sample in comparison to other samples. You will visualize relative variation instead of absolute variation. To find out if this is something you would like, you can consider the following example:

In Sample 4, which population is most extreme? Is it population A, because the difference to the mean is largest there? Or is it population B, because it deviates more from the other samples? If your answer is population A you are interested in absolute differences and you should use simply Mean = 0. If your answer is population B you are interested in relative differences and you should use Mean = 0, Var = 1.

- Using Var = 1 means that all features (variables) have equal impact on the result, irrespective of whether they vary a lot between samples or not. This might be desirable in some cases, but undesirable in others. If you have many features which are non-informative, i.e. roughly constant across samples, using Var = 1 should generally be avoided.

LOG TRANSFORMATION

Log transformation can be selected in the **Data** tab. It impacts all visualizations and statistical tests.

Using a log transformation means that deviations between samples will be measured by fold change instead of by direct differences (subtraction) of values. This means for example that a frequency change from 1% to 2% will be equivalent to a change from 10% to 20%.

When using a log transformation, you need to threshold your data to a minimal positive value, since you cannot define a fold change with negative or zero values. The threshold should be chosen such that the minimal value reflects a baseline from which you measure fold change. Variations below the baseline will be disregarded. For example, if you use MFI's you can use a typical MFI for a negative population as threshold.

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.

REFERENCES

N. Aghaeepour et al. (2013). Critical assessment of automated flow cytometry data analysis techniques, Nat Methods 10, 228-238.

J. Spidlen et al. (2012). FlowRepository - A Resource of Annotated Flow Cytometry Datasets Associated with Peer-reviewed Publications. Cytometry Part A 8, 727-731.