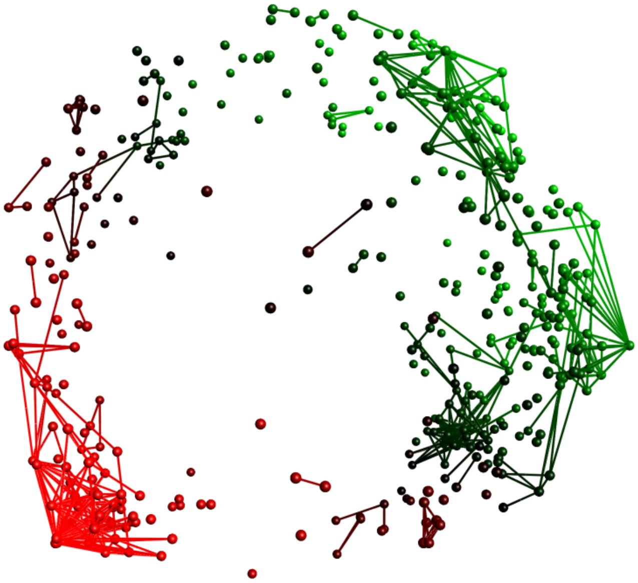
Qlucore Omics Explorer Tutorial

An Exploration and Visualization Tool



Analysis in an instant

Contents

Preface

Qlucore Omics Explorer is supplying new technology in data analysis and data mining. It is built on state-of-the-art mathematical and statistical methods.

The main features of Qlucore Omics Explorer are the ease and speed with which you will be able to analyze and explore your data sets. You will rapidly and easily visualize and work interactively with your data sets in real time directly on the computer screen. Throughout the analysis you are supported by general statistical methods. Data import and export is easy with a wide range of options.
You are not expected to have in-depth knowledge of mathematical or statistical methods. With an ordinary computer you will be able to easily explore your high-dimensional data and rapidly find relevant results. All results will be presented to you instantaneously.

This tutorial is meant to enable first-time users to understand and use the basic capabilities and features of Qlucore Omics Explorer. A comprehensive description of all functions can be found in the **Reference Manual** that is supplied in the **Help Menu** of Qlucore Omics Explorer. More information is provided on www.qlucore.com/documentation.aspx. There you can also watch instruction films.

Welcome to an interactive and explorative journey!

Overview

This tutorial is divided into four main sections.

1. Import data or load example files
2. Familiarize yourself with the user interface
3. Use visual interaction and perform a statistical test
4. Explore a data set and be guided to several areas of key functionality

It is recommended that you work through all the 4 steps of the tutorial.

In the “Appendix: An introduction to the statistical concepts” on page 58, you can read about the basic statistical concepts used in Qlucore Omics Explorer.

Recommendation:

From the Getting Started dialog (visible from the start or available from the **Help** menu) select to watch the Video tutorials as a complement to this written tutorial.

Data and annotations import and example files

Qlucore supports a number of import methods and workflows. There are three main methods of data import.

1. Import and normalization of platform dependent "raw" data. Qlucore Omics Explorer supports for instance Affymetrix gene expression arrays, Agilent mRNA and microRNA arrays and RNA-seq data from BAM files.
2. Platform independent import of normalized data from a text based file (.txt, .csv and .tsv).
3. Direct download from the Gene Expression Omnibus (GEO).

Data can be generated from many sources and the list below presents a number of examples:

- Gene expression: microarrays, real-time PCR, RNA-seq
- MicroRNA: microarrays, real-time PCR, RNA-seq
- DNA methylation: microarrays
- Protein expression: microarrays, antibody arrays, 2-D gels, LC-MS data
- Flow cytometry
- Proteomics

To get the latest information on data import, check the document "How to import data", which can be found on www.qlucore.com/documentation.aspx.

Annotations

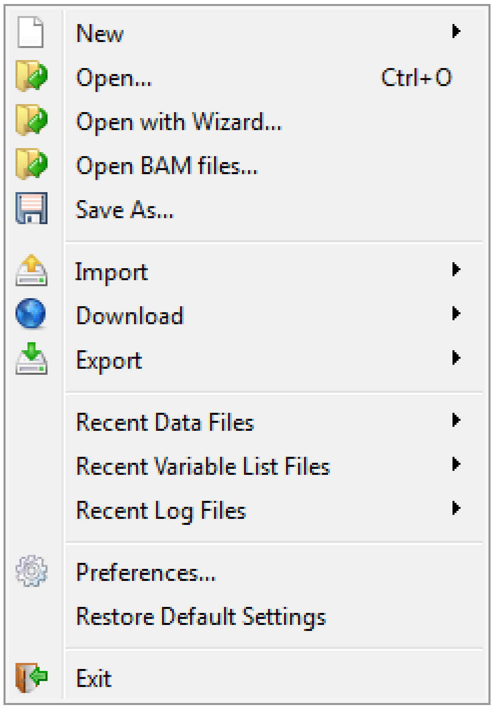
In addition to importing data you normally also want to import annotations[1]. In OE we use the terminology sample annotations for information about the samples such as treated/non-treated, age or gender. Variable annotations are defined as information about the variables such as weight, gene symbol or chromosomal location. Basic variable annotations are normally provided by the manufacturer of the instrument that you have used to generate your data[2]. Sample annotations you normally create yourself, and you may have to import them separately or create and edit them manually in the program.

Open a file (data or annotations)

- Use the **File->Open** command to open a data file.

[1] Variable annotations are imported automatically if you use an Affymetrix array.

[2] If you use the import wizard for normalized data you have the possibility to import sample annotations, variable annotations and data, at the same time.



Depending on the extension of your file QOE will select an import method. For non specific extensions (i.e. ".txt", ".csv", ".tsv") you will be guided to the appropriate import method through a suite of questions.

If you have data from other platforms than Agilent and Affymetrix it is normally enough to save a normalized version of your data into a tab separated text file. Then select the option **Open with Wizard**. Most instrument manufacturers provide the necessary software to do this as part of their instrument. The file can include both measurement data and annotations. When you open a file like this the Wizard will help you to select which parts of the file to import.

If you have BAM files with RNAseq data to import, select **Open BAM files**.

- To import sample annotations first open a data set and then select **Import > Sample annotations** and for variable annotations select **Import > Variable annotations**.

Direct download of data from GEO

QOE provides fast and easy direct download of all datasets from NCBIs Gene Expression Omnibus (GEO) https://www.ncbi.nlm.nih.gov/geo/

You can browse all the available datasets using GEO's own browser tool https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

Once you have chosen which dataset that you would like to explore using QOE, you simply select **File>Download>GEO Data Set** in QOE.

Example files

The following two example data files are supplied with the program.

Qlucore Test Data Set.gedata
This is a synthetic data set. It includes 12 samples and 50 variables.

Acute Lymphoblastic Leukemia.gedata
This data set consists of gene expression profiles from 132 different patients, all suffering from some type of *pediatric acute lymphoblastic leukemia (*ALL). For each patient the expression level of 22282 genes has been measured. The data set comes from a study by Ross et. al. [Ross2003].

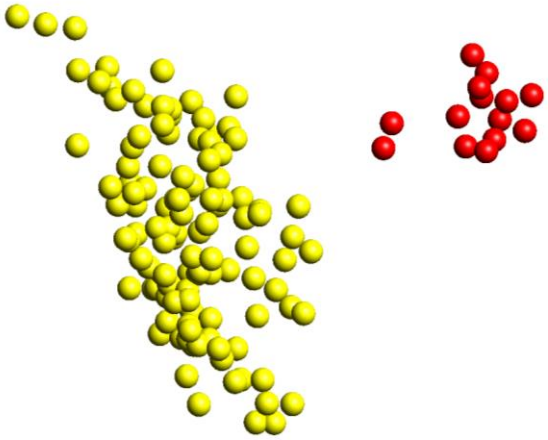
In addition to the example data files, the program includes two example gene set collections (available from the **GSEA Workbench**) as well as a slim Gene Ontology (available in the **GO Browser**).

How to interpret a PCA plot

The basic meaning of the PCA plot of any multi-dimensional data in QOE is that data points that are similar based on the observed data values are also presented close together in the generated plots.

The PCA operation reduces the dimensionality of the data in order to generate three-dimensional graphical representations. In this process, as much as possible of the information (the variance) in the original data is retained.

In the picture below, the red group consists of data samples that are similar to each other and that are different from the samples in the yellow group.



Start to use Qlucore Omics Explorer (QOE)

To start QOE, double click the **Qlucore Omics Explorer icon** (shortcut) on the desktop or start **Qlucore Omics Explorer** from the **Program Menu** found in the **Start** Menu.

The QOE **Main Window** will appear. We begin with a quick orientation of what you see on the screen, see the figure below. In the middle you find the **Work Space** where all plots will be displayed in **Plot Windows**. Furthermore you find several **Dock Windows**. By default the **Samples, Variables** and **Log** dock windows are docked to the left of the main window and the **Statistics** and **Getting Started** windows are floating.[3]

In the **Menu Bar** you can select the displayed dock windows under **View> Dock Windows**. From the **View** menu you can also launch the **GO Browser** and the **GSEA Workbench** as well as the Quality Control (**QC**) report. Under the Menu bar you find different controls that govern the functionality of the mouse tool and the operations performed on the data set. You also find seven different **Tabs**: **Data, Method, Options**, **View, Cluster, Build classifier** and **Classify**, that will help you to select and manage the work flow in QOE.

In the **Statistics dock window** you can select which statistical methods that you want to use to study your dataset.

Tip: If you have closed the Statistics dock window go to the View->Dock Windows menu and make it visible again.

Finally at the bottom you find the **Status Bar**. In the Status Bar you find for instance the total number of samples and variables in your data set displayed and information on how many of them that actively takes part in the analysis at the moment.

To begin, we first of all restore default settings.[4]

- Select the **File** > **Restore Default Settings** menu item in the **Menu Bar**
- Select **OK**, when you are asked if you want to restore default settings.

Note that when you later exit QOE the current settings will be saved and used the next time the program is started and you can thus directly start working with the settings that best suit your data.[5]

[3] You can rearrange the dock windows any way you like: docked to the left or the right, floating or hidden. You do this by grabbing and moving their title bars with the mouse functionality.

[4] This is not something you have to do every time. We do it here in the example to avoid possible discrepancies between used settings.

[5] Thus in order to have the default settings restored you will have to exit and then restart OE.

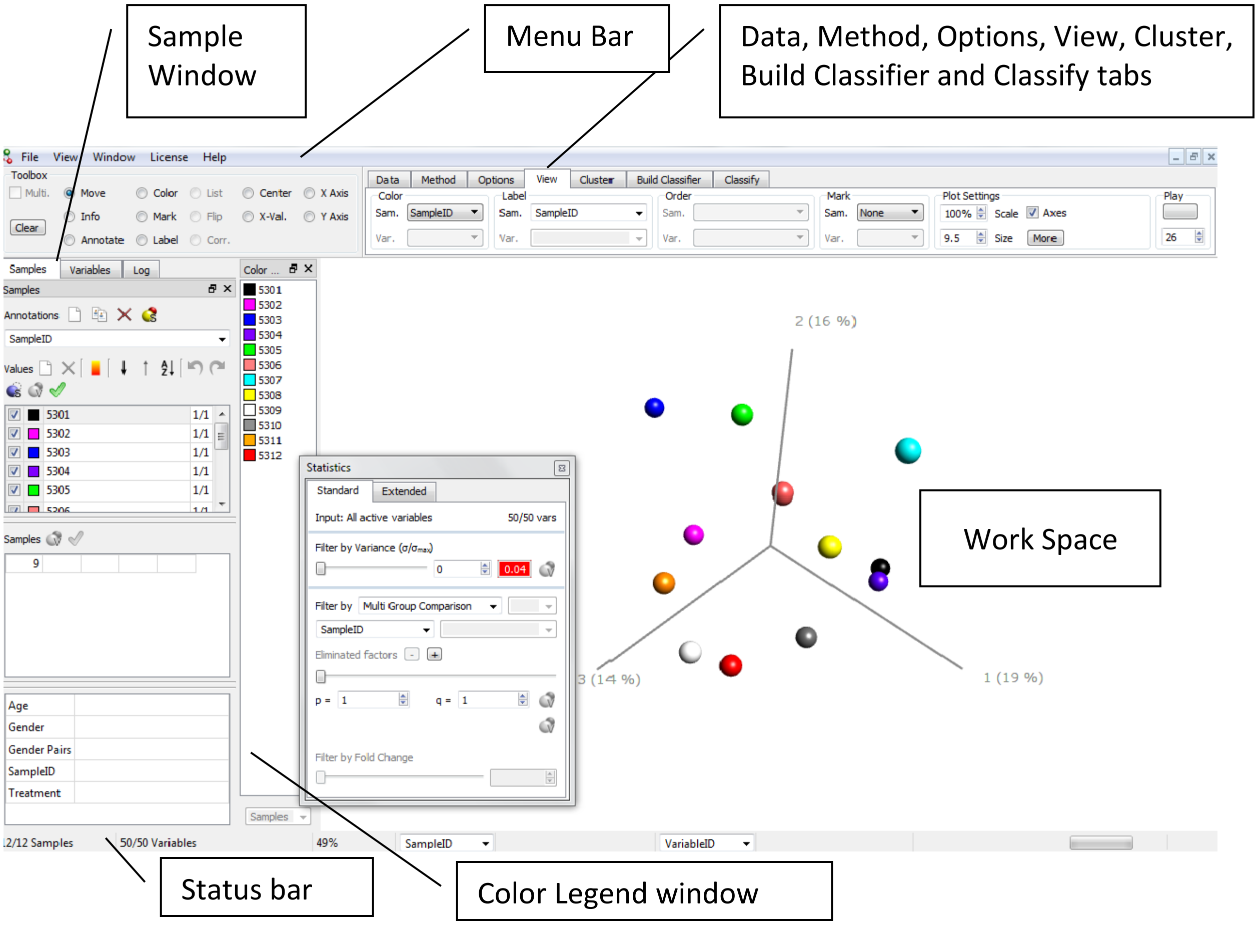
We are now ready to open a data file.

- Go to the menu **Help> Example Files> Qlucore Test Data Set.gedata**[6]

Tip: To close a data set close the last open plot for the data set.

The **Qlucore Test Data Set** is now open in **QOE** and you have the starting position for beginning to analyze the data.

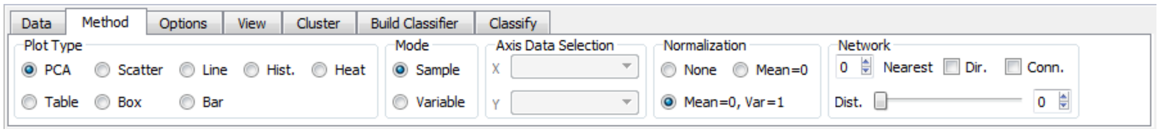
What you presently see in the **Work Space** is a **principal component projection** of the 12 **samples** from 50 dimensions (corresponding to the 50 measurements (variables) for each sample) down to the three-dimensional space spanned by the three first principal components. The samples are colored according to the annotation sample ID as can be seen in the **Color Legend** window.



[6] Usually when you want to open a file you use File>Open and then you select the file you want to open by double clicking on it.

To the left, the **Sample window** is selected by default. Here you can see and manipulate information connected to your samples. Correspondingly, you will find information relating to the variables by selecting the **Variable window**. By selecting the **Log window** you will be able to create a Log of your workflow in QOE. In the **Color Legend window** you find all relevant information, in a form which is easy to export, concerning coloring of samples or variables. You can choose to close the Color Legend dock window in order to get more work space.

Above the Work Space and under the menu bar you find different controls containing commands for some of the basic functionality. Among them are the **Data, Method, Options** and **View tabs**.
The **Data tab** has controls related to how the data is prepared for future analysis and the **View tab** contains controls related to how the data is presented.
The Method tab is selected by default and there you find controls related to analysis.
The **Options tab** includes refinements and additions of the methods available in the **Method tab**. You select the **plot type** in the **Method tab** (**Sample PCA** is selected by default). You also select how to **normalize**/**scale** your data here and to **create graphs/networks**.
The **Cluster tab** controls the automatic creation of clusters.
The **Build classifier tab** includes functionality to build classifiers.
The **Classify tab** enables classification of samples based on a classifier.



There are several different plot types available in QOE and it is possible to configure the plots in various ways. There are eight main plot types and you select the plot type in the **Method** tab.

- PCA
- Heatmap (Heat)
- Scatter
- Table
- Line
- Box
- Bar
- Histogram (Hist.)

Several of the plots can be configured to show either samples or variables. This is selected using the **Mode** buttons in the **Method** tab.

In **Scatter**, **Box**, **Line, Bar** and **Histogram** further selections on what to plot is made using the **Axis Data Selection** tools. We will present some of the options and how to configure a plot using these tools later in the tutorial, see the section “Further analysis” on page 50.

QOE allows a high degree of flexibility in the number of plots and datasets that you can have open in parallel. When many plots are open the plot that last was activated by left clicking with the Mouse is called the **Active plot**. Selected operations will normally affect the **Active plot.**

The **Statistics window** is important and you have the freedom to position this window wherever you want on your screen. We will later describe some of the functionality of the Statistics window.

- Move the **Statistics window** to clearly see the displayed samples.

Before we continue we will familiarize ourselves with the basic structure of **the Qlucore Test Data Set.** A subset of the data set is presented in the table below. The data set includes 50 variables measured for 12 samples.

Sample annotations

Variable annotations

			SampleID	5303	5308	5302
			Age	20	26	28
			Gender	Female	Female	Male
			Gender pairs	P3	P1	P3
			Treatment	Drug 2	Placebo	Drug 2
VariableID	**Symbol**	**Name**				
ID_01	ENC1	Variable 01		0.071	0.022	-0.027
ID_02	CDK8	Variable 02		1.541	-0.316	0.508
ID_03	PEX7	Variable 03		0.205	0.234	-0.279
ID_04	VPS13D	Variable 04		0.635	1.965	0.781
ID_05	DLX6	Variable 05		0.046	0.230	-0.127
ID_06	SFRS9	Variable 06		-0.245	0.082	-0.079
ID_07	ABCF1	Variable 07		0.077	0.559	0.139
ID_08	DAD1	Variable 08		-0.673	-0.017	0.062

The first *five* rows are the sample annotations and the first *three* columns are the variable annotations.

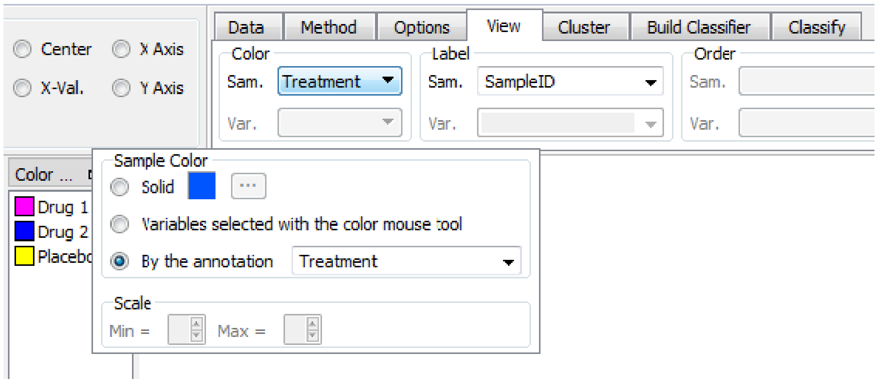
The data matrix starts with the cell with the value “0.071”.

Now, let’s continue.

All commands given in QOE immediately affect the projections displayed in the Plot Windows.

- Select the **View tab**

In the **Color** section of the **View** tab, select to color your samples according to the annotation “Treatment”.



Color: Sam. Treatment

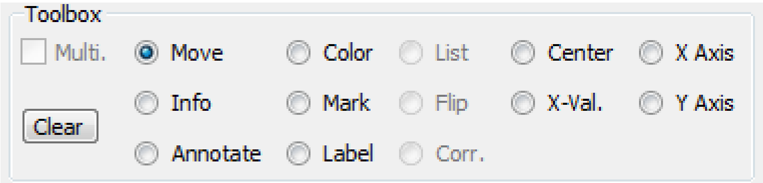
Label: Sam. SampleID

🟪 Drug 1
🟦 Drug 2
🟨 Placebo

Sample Color: ● By the annotation Treatment

This will color the active **PCA sample** plot. In the **View tab** there are many more coloring options and the options will change depending on the type of plot that is active. In the tutorial we will mention several of the coloring options but do not hesitate to try on your own.

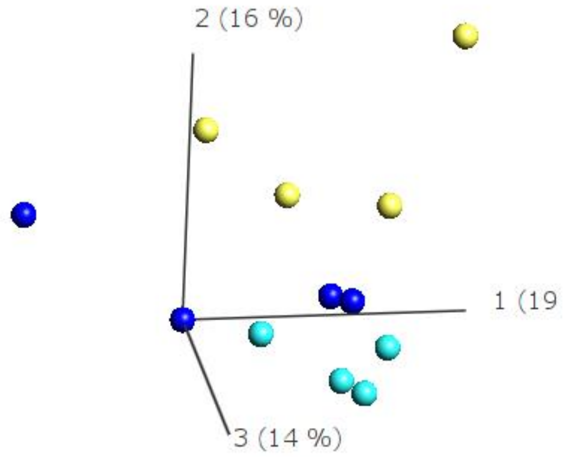
In the **Tool Box** you select the function of the mouse (the radio button **Move** in the upper left corner is checked by default).



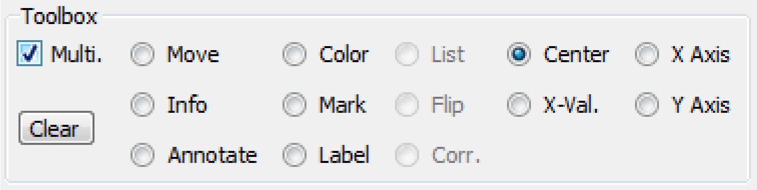
● Move

When **Move** is selected, you can rotate the image by holding down the left mouse button and *drag the image* with the mouse in the Work Space.

- Select **Center** in the **Tool Box** and then **left-click on a sample in** the plot**.** The selected sample is then placed in the center of the **Plot Window**.



By left-clicking another sample, this sample will be placed in the center instead. By clicking **Clear** in the **Tool Box** the original plot is restored.



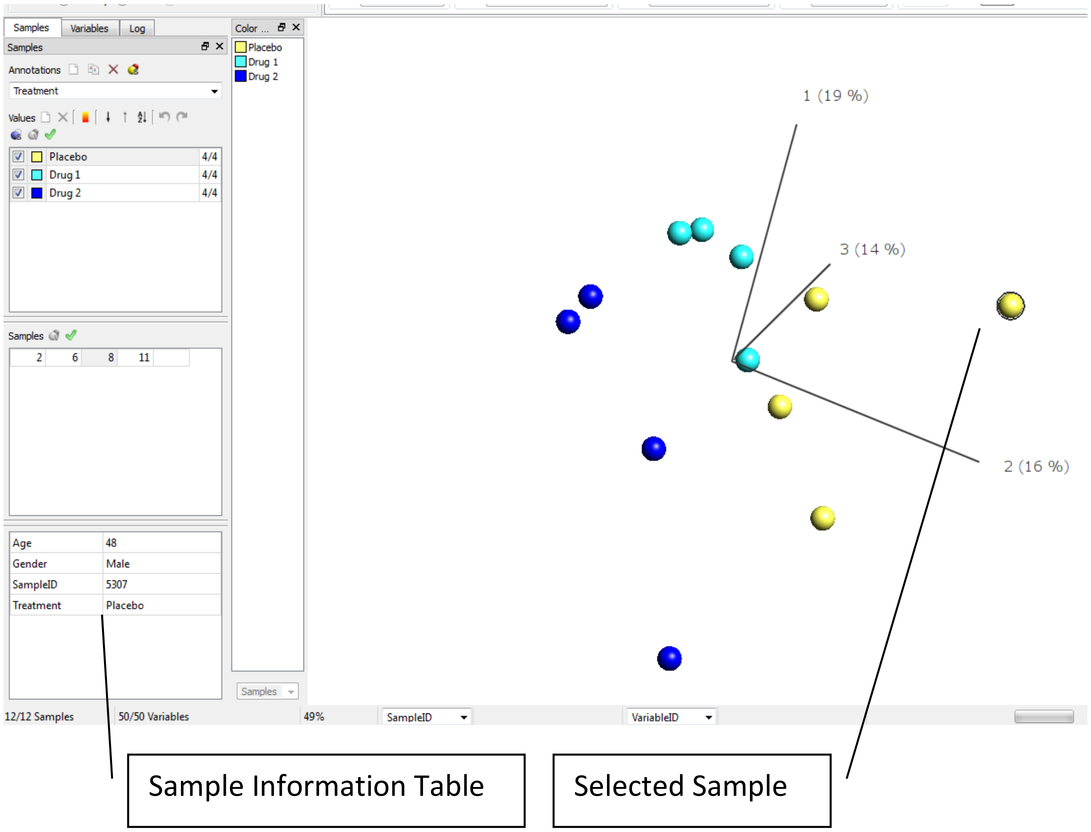
Clear and **Multi** are two controls that affect the use of the selected mouse tool. **Clear** will remove all selected marks and labels and **Multi** will allow you to make multiple selections.

- Select **Move**.

In the lower left corner of the Main Window you find the **Status Bar**. Here you see the text **12/12 Samples**, indicating that all available samples are currently considered in the workflow, each one of them corresponding to one of the 12 small spheres you see plotted on the screen.

- Choose **Info** in the **Tool Box** and then left click on a sample.

You then get the available annotations for the particular data frame corresponding to that sample in the **Sample Information Table.**



☑ 🟨	Placebo	4/4
☑ 🟦	Drug 1	4/4
☑ 🟦	Drug 2	4/4

Samples: 2 6 8 11

Age	48
Gender	Male
SampleID	5307
Treatment	Placebo

🟨 Placebo
🟦 Drug 1
🟦 Drug 2

12/12 Samples 50/50 Variables 49% SampleID VariableID

Sample Information Table

Selected Sample

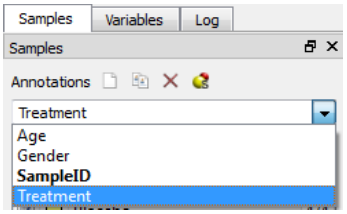
Note: *What you see on your screen can differ from the images by a rotation, due to the fact that we have been practicing rotation above. This should not cause confusion and in fact we recommend you to continuously rotate the images during the work-through of the example to get a good feeling for the three-dimensional structure of the projection. When **Move** is selected in the Tool Box you can always return to the starting position by selecting **Clear** in the Tool Box.*

Sample information

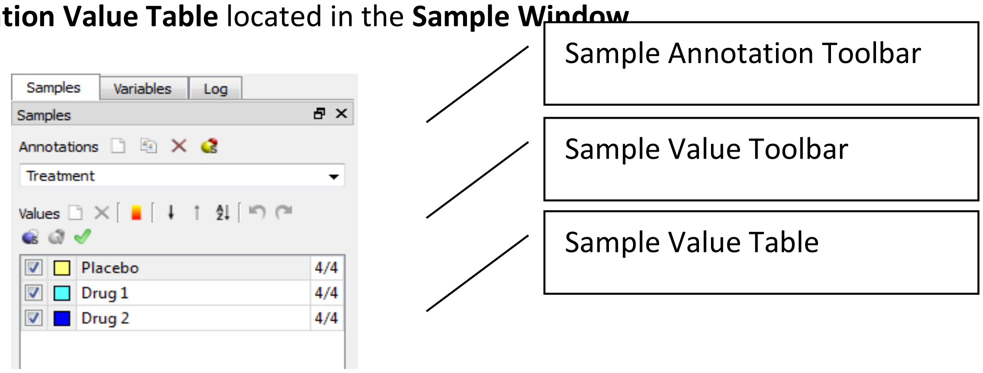
We shall now see how sample information (annotations) are presented and used.

In the **Annotation combo box** located in the **Sample Window** you can choose between which of the different annotations that come with the data frame that you want to work with.

- Select “Treatment” from the list of **Sample annotations**



This makes the different treatments used in the study appear in the **Sample Annotation Value Table** located in the **Sample Window**.



Sample Annotation Toolbar

Sample Value Toolbar

Sample Value Table

In the **Sample Value Toolbar** and in the **Sample Annotation Toolbar** you see some of the (buttons) icons present in QOE.

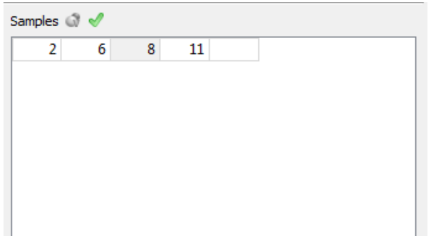
The **Sample Color** buttons S give an instruction to color samples in the active plot according to the annotation chosen and the **Variable Color** buttons V give an instruction to color variables.

The **Mark** buttons ✓ give an instruction to mark a selected sample group or group of variables.

The Sample Color, Variable Color and Mark **buttons are short commands** for functionality also available in the **View tab**.

Note that the name of the function of a button appears if you leave the mouse tool over it for an instant.

In the **Value Table** you see the three different subtypes of "Treatment" listed.
By selecting the different rows in the **Value table** you see the corresponding sample numbers displayed in the **Sample Window** below the **Value table.** By choosing for instance the **Placebo** subtype you get the following Sample Window listing the 4 samples in this subtype group (Placebo).

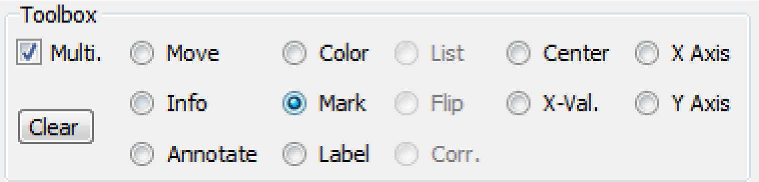


2	6	8	11

Note: The sample numbers are automatically created by QOE when a file is opened. The main functionality of the numbers is to provide a unique identification for each sample.

You choose the functionality of the mouse tool in the **Tool Box**.

- Select **Mark** in the Tool Box



Toolbox

☑ Multi. ○ Move ○ Color ○ List ○ Center ○ X Axis

○ Info ● Mark ○ Flip ○ X-Val. ○ Y Axis

Clear ○ Annotate ○ Label ○ Corr.

Note that **Multi** is selected by default, thus making it possible to mark multiple samples.
By **left clicking** with the mouse directly on a sample in the plot, or in the **Sample Annotation Table** or on a Sample subgroup in the Sample Value Table, the corresponding Samples are marked.

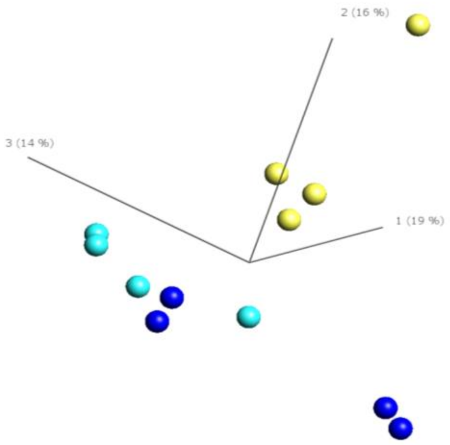
If you would like to select many objects at the same time, you can also circle clock-wise, in a plot, all of the objects using the mouse pointer. If you circle counter clock-wise you will remove the objects from the selection.[7]

The same functionality applies to the other options (e.g. **Color**, **Label** and so on) in the **Tool Box**. When you select the functionality of the mouse tool in the **Tool Box** you can afterwards make selections with the mouse tool in every table, plot or window in QOE where the selected functionality is relevant and meaningful.

- Select **Clear** to clear all marks.
- Select **Move** in the Tool Box Window to be able to freely rotate the plots.

You now see the different subtypes of “Treatment” colored according to subtype in the **Work Space**.

[7] The classify tool works by adding samples to a group and hence counter clock-wise selection is not defined.



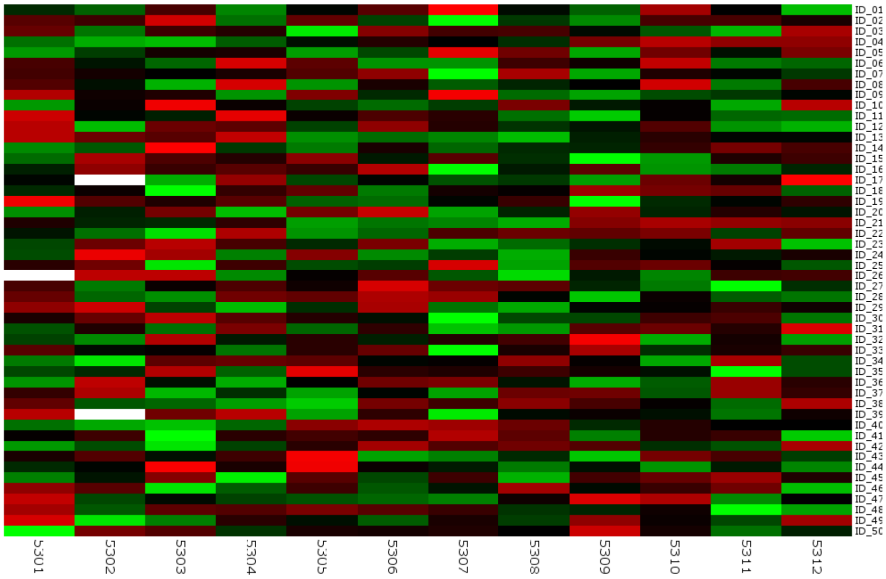
At this point we have guided you in the basic interaction method of QOE using the **Mouse tool** and the **View tab**. You have generated and colored a sample PCA plot. There is a lot more basic functionality and we will go through it step by step. However, it is now time to combine visualization with statistics and demonstrate how to do a statistical test.

Statistical Analysis with visual feedback

In this section we will continue to work with only one active plot and it will be a heatmap. We will also introduce the statistics functionality. In the “Appendix: An introduction to the statistical concepts” on page 58, you can read about the statistical concepts used.

We will do a statistical test and visualize the result simultaneously. We will use the heatmap to generate the visual feedback. We could have worked with any plot type but for pedagogic reasons we introduce a new plot type. Select **Heat** in the **Method** tab.

You will directly get the plot below.

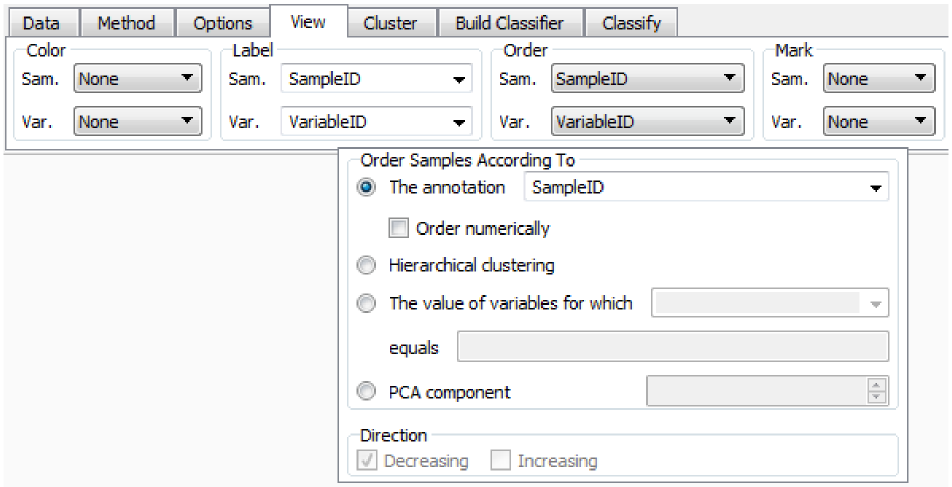


Note: The white parts of the heatmap indicate values that have been reconstructed using missing value reconstruction. In the ***Data tab*** *the method for missing value reconstruction can be chosen. Please refer to the* ***Reference*** *manual for more details about missing values.*

You select the Normalization in the **Normalization box** in the **Method tab.** The default is to present normalized data.

First we enhance the plot by adding additional visual elements.

- Select **Color Samples** in the **View tab** and then color by the annotation *"Treatment"*.
- Then select **Order Samples** in the **View tab** and select **"Hierarchical clustering"**



Color
Sam. None
Var. None

Label
Sam. SampleID
Var. VariableID

Order
Sam. SampleID
Var. VariableID

Mark
Sam. None
Var. None

Order Samples According To
● The annotation SampleID
□ Order numerically
○ Hierarchical clustering
○ The value of variables for which
equals
○ PCA component

Direction
☑ Decreasing □ Increasing

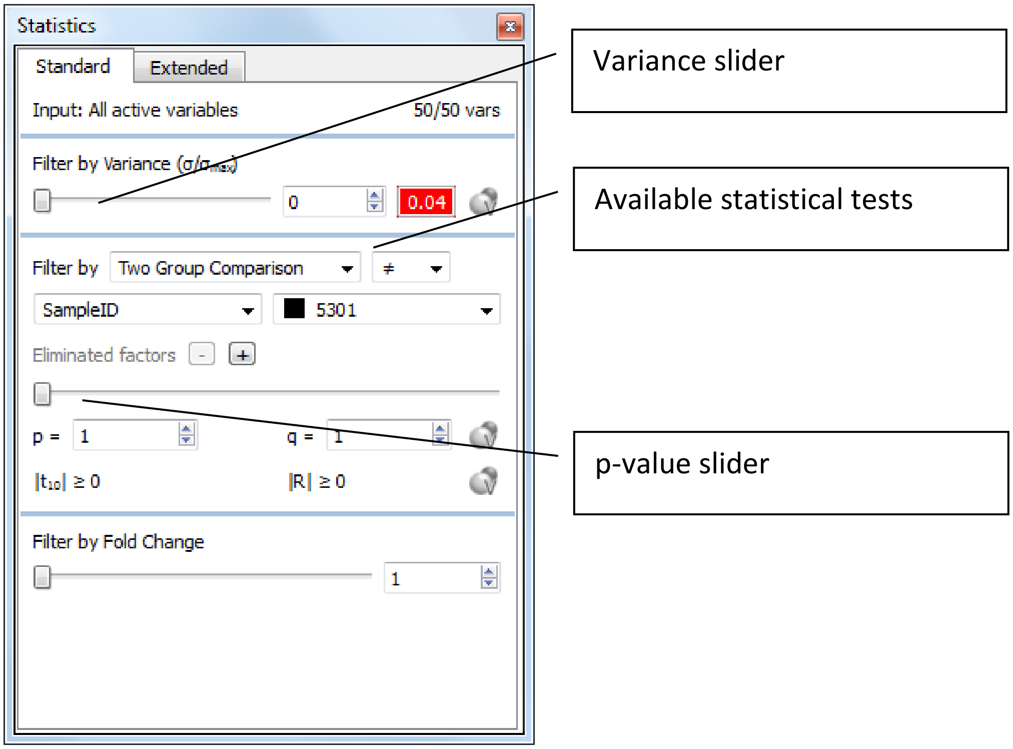
There are four different algorithms (**Linkage**) that you will find in the **Options tab** for generating the clusters (mean, weighted mean, minimum and maximum linkage). We refer to the reference manual for more information on this and related information on heatmaps and clustering. The clustering can be based either on

- Covariance (i.e. using data normalized to mean 0 for each variable), or
- Correlation (i.e. using data normalized to mean 0 and variance 1 for each variable).

We are now ready to introduce a statistical test to find the variables that best separate "Placebo" from "Drug 1" and "Drug 2".

In the **Statistics Dock Window** you control what type of statistical test you would like to set up. The available tests are:

- Two Group Comparison (t-test)
- Multi Group Comparison (F-test)
- Linear Regression
- Quadratic Regression
- Rank Regression



Input: All active variables 50/50 vars

Filter by Variance (σ/σ_{max}) 0 0.04

Filter by Two Group Comparison ≠

SampleID 5301

Eliminated factors - +

p = 1 q = 1

$|t_{10}| \geq 0$ $|R| \geq 0$

Filter by Fold Change 1

The other components of the statistical dialog are the **Variance slider** at the top, the **p-value slider** and the **Fold Change slider**. The use of the variance slider is covered in section "Basic exploration" on page 27. The **p-value slider** is used to select the p-value cut off for the chosen statistical test. The Fold Change is applied after the Variance and the statistical tests. Fold change is only defined for two group comparisons.

At the top of the statistics dialog is the input information. It shows how many variables that are used as input to the filters. If you have not done any selections in the Variable tab it should say “All active variables”, more about this on page 37 and onwards.

The red box is the Projection score, more about this in section “Basic exploration” on page 27.

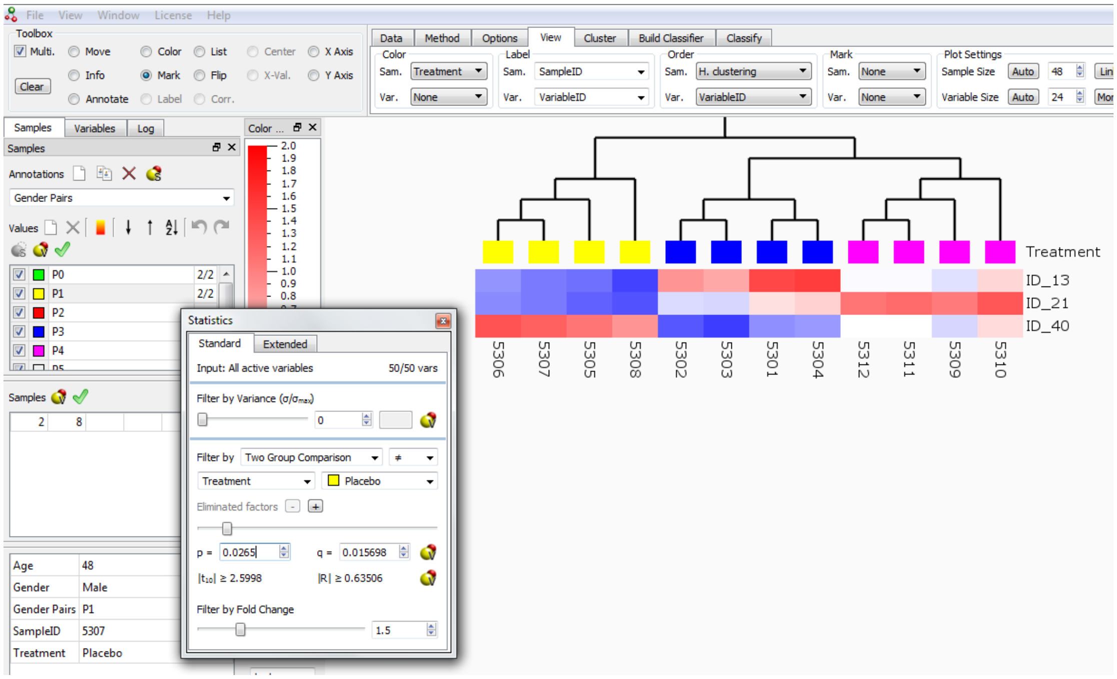
Note: The Advanced tab is the interface to statistical r-scripts. See the reference manual for more information.

To set up the test that finds the variables that best separate “Placebo” from “Drug 1” and “Drug 2”, and have a Fold change of at least 1.5, carry out the following steps.

- Select to Filter by **Two Group Comparison** (t-test) in the Statistics window.
- Select “Treatment” in the corresponding Combo box
- Verify that “Placebo” is highlighted in the third Combo box
- Move the Fold Change slider to 1,5.

Adjust the **p-value slider** to a p-value of 0.0265, do it slowly. You can see how the heatmap is updated continuously and the number of variables that fulfill the cut off criteria is decreasing. There are 3 variables that have a p-value of 0.0265 or lower for the selected t-test (Testing “Placebo” against the samples in the groups “Drug 1” and “Drug 2”) and have a Fold Change of at least 1.5.

The corresponding q-value (0.015698) (which can be interpreted as a false discovery rate) is also displayed in the **Statistics** window and the 3 variables that now are left in the analysis have a high statistical significance. How much you should filter depends on the structure of your data and what significance levels you want to achieve. The results are shown in the plot below.



View tab — Color: Sam. Treatment, Var. None; Label: Sam. SampleID, Var. VariableID; Order: Sam. H. clustering, Var. VariableID; Mark: Sam. None, Var. None; Plot Settings: Sample Size Auto 48, Variable Size Auto 24

Annotations: Gender Pairs

☑ 🟩 P0	2/2
☑ 🟨 P1	2/2
☑ 🟥 P2	
☑ 🟦 P3	
☑ 🟪 P4	

Samples: 2 8

Age	48
Gender	Male
Gender Pairs	P1
SampleID	5307
Treatment	Placebo

Statistics — Input: All active variables 50/50 vars

Filter by Variance (σ/σ_{max}) 0

Filter by Two Group Comparison ≠

Treatment — Placebo

Eliminated factors

p = 0.0265 q = 0.015698

$|t_{10}| \geq 2.5998$ $|R| \geq 0.63506$

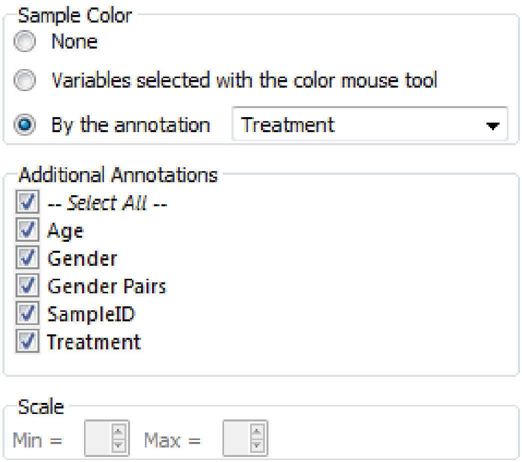
Filter by Fold Change 1.5

Using the plot we can do multiple observations.

- The three variables (ID_13, ID_21 and ID_40) have the best statistical significance (since they are present) and they also have a Fold Change of at least 1.5.
- The first branch (counting from the top) of the Hierarchical clustering divides the Samples into two clusters; "Placebo" and "Drug 1 and Drug2". The second branch splits the "Drug 1 and Drug 2" cluster into "Drug 1" and "Drug 2". This is expected since we have used the statistical test to identify variables that actually are best at separating "Placebo" from "Drug 1 and Drug 2".
- From the regulation we can for example conclude that the variable ID_21 has high values for samples in the group "Drug 1" and that the variable ID_13 has low values for all samples in the "Placebo" group. With default settings, red color in the heatmap indicates that a variable has a high value for that sample and green indicates that a variable has a low value for that sample.[8]
- Select the **Label box** in the **View tab**. Label the variables according to the variable annotation "Symbol".
- Select the **View tab** and **Order** the variables according to **Fold Change**
- Select the **View tab** and **Color** variables according to **Fold Change**.
- Select the **View tab** and Color samples for all annotations. "Select all"

A natural next step is now to find out more about the variables that we observe in the plot. First we shall label the variables using a new annotation.

[8] The color scale can be adjusted. Select Plot settings in the View tab and then the More button.



Sample Color

○ None

○ Variables selected with the color mouse tool

● By the annotation Treatment

Additional Annotations

☑ -- *Select All* --

☑ Age

☑ Gender

☑ Gender Pairs

☑ SampleID

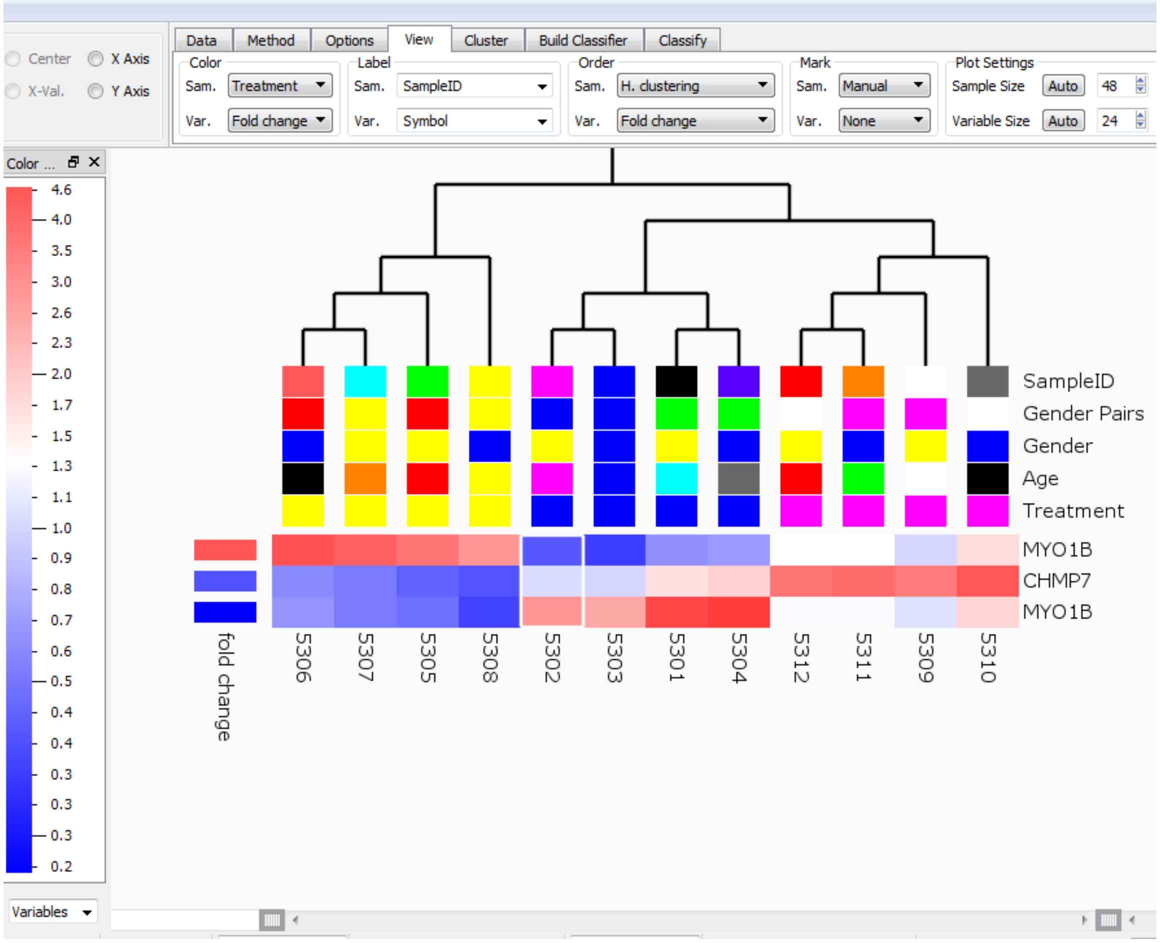
☑ Treatment

Scale

Min = Max =

One of the key usage models in QOE is that it shall be very easy to change the analysis path. To give an example on this select the **Sample window** and select the “Treatment“ annotation. Then deselect the “Drug 2” group check box. This removes all samples for which the "Treatment" annotation equals "Drug 2", and updates the results accordingly. The updated result is immediately presented, see the picture below. The result is an updated statistical test which finds the variables that best separates “Placebo” from “Drug 1”. *With the same settings in the statistics dialog more variables (4) are now found that match the earlier selected test criteria.*

This will give you the following plot. It contains a lot of information. The different ways to change coloring, labeling and ordering provides extensive flexibility to tailor the plot to your needs.



Color: Sam. Treatment; Var. Fold change
Label: Sam. SampleID; Var. Symbol
Order: Sam. H. clustering; Var. Fold change
Mark: Sam. Manual; Var. None
Sample Size 48
Variable Size 24

The **Color Legend** shows the color scale for the variables which are colored according to **Fold Change.** As an example we observe that the variable with the "Symbol" name MYO1B has the highest **Fold Change**.[9] Also observe since we have been working with a two sided test the Fold Change setting of 1,5 applies to both directions, i.e. +1,5 and - 1,5.

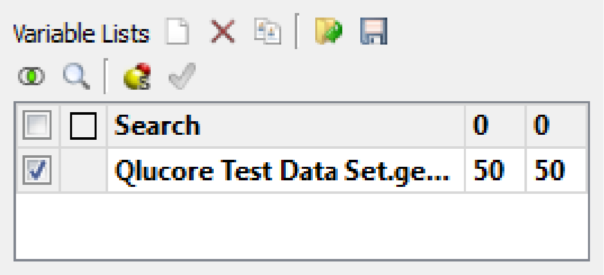
There are many plot types that now can be used to provide deeper insights into the findings such as Kaplan-Meier if the data include survival information, box plots for providing detailed information about how variables vary of different subgroups and the bar plot to visualize data after multiple annotations such as time and treatment. This will be covered in the section "Further analysis and exploration" on page 50.

There are many ways to save the obtained results. The **Log** functionality in the **Log Dock Window** can be used, another option is to select **File > Export > Image** to export

[9] Observe that to get correct levels for the Fold Change you need to work with logarithm-transformed data. The logarithm can be taken in OE or before importing data. Please refer to the Reference Manual for more details.

the plot. It is also possible to save the list of variables, which is what we will demonstrate below

- Select the **Variable Window**.



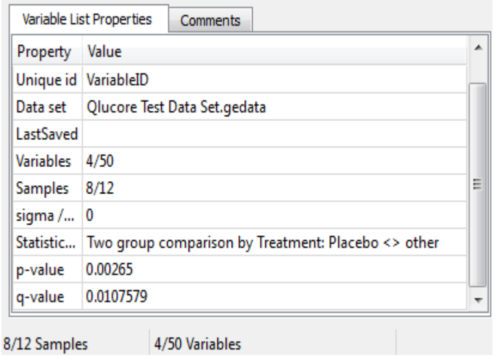
☐	□	Search	0	0
☑		Qlucore Test Data Set.ge...	50	50

The **Search** list is special and it is created when the first data set is loaded. The **Search** list includes the result of the last search done. A search is initialized by pressing the 🔍 button. More information about this in the section: ”Working with Variables”, on page 37.

The second list is the **Active list.** There is one **Active list** for each open data set. The **Active list** has the same name as the data set. In the **Variable List Table** all open variable lists are listed. In addition to the **Search list** and the **Active list** this table can contain lists that are imported manually, or lists generated with the **GO browser** or exported from the **GSEA Workbench.**

The **Active list** includes all the active variables in a data set. Now the list includes 4 variables that match the performed statistical test. Change the **p-value slider** in the **Statistics Window** and observe how the list is updated to include fewer or more variables. Before saving the list, it is useful to populate it with relevant information.

- Make a copy, 📋 of the **Active list.**
- Use the **Select Columns** button ▶ and add columns of interest. Select “p-value”, “q-value” and “Symbol” from the list. These columns are then added to the list.
- In the Variable List information part of the Variable tab there are two information elements: the **Variable list properties** and the **Comments**. You can add your own information about the list in the Comments field. The Variable list properties fields will include information of how the list was created. In the properties below we can for instance see that the list was created using a Two group comparison and that 8 samples out of 12 were active.



Property	Value
Unique id	VariableID
Data set	Qlucore Test Data Set.gedata
LastSaved	
Variables	4/50
Samples	8/12
sigma /...	0
Statistic...	Two group comparison by Treatment: Placebo <> other
p-value	0.00265
q-value	0.0107579

8/12 Samples　4/50 Variables

- Select the 💾 button, enter the file name and location to save the file. The next step is to decide what to include in the list. In the Export Variable List dialog you will be presented with several options, such as whether to save only the Active variables or all variables, and whether to include annotations or not.

Note: The values of the statistical calculations for a variable in a variable list are updated dynamically when the input is changed.

The variable list is now saved and it includes 4 rows with information as the Variable ID, the “Symbol”, p-value and q-value. The file is a tab separated text file with information about how the list was created and you can open it using any Spreadsheet program or a normal text editor. If you need a plain text file without any comments use the “Plain text file format” option in the **Save** dialog.

Tip: Try a couple of the different options and open the saved file in a spreadsheet program or an editor and examine the results.

In this section we have done a statistical test (t-test + Fold Change filtering) and visualized the results using a Heatmap. During the analysis, one sample group was deselected and the work continued on a subset of the data. Finally the result was saved in a variable list.

Data set exploration

In this section we will demonstrate how to approach a new data set and how to obtain understanding and information about the data set including using system biology information.

Basic exploration

QOE is very well suited to investigate and explore a data set. This type of work is often also called data mining or hypothesis generation. The user is looking for both expected and completely new things in a dataset and speed and interactivity are essential components to support this type of work.

For this part the Acute Lymphoblastic Leukemia data set will be used as the example.

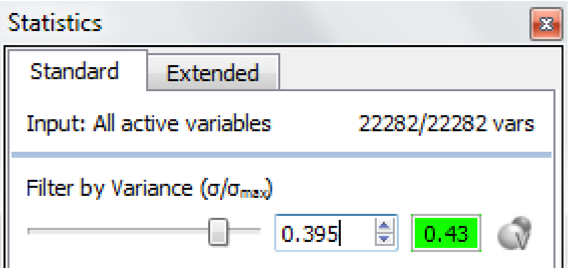
- Open the Acute Lymphoblastic Leukemia data set by **selecting Help -> Example Files -> Acute Lymphoblastic Leukemia.gedata**.

Even when we rotate this picture (select the **Move** functionality in the **Toolbox**) it is very hard to discern any structure or pattern in the plot. The reason for this is that all 22,282 genes (variables) take part in the analysis. Most of them have possibly very little to do with the different genetic variations that we are interested in, but all of them contribute to the noise in the data by small random fluctuations.

At this point the projection score will support us in the analysis, see the figure below. The projection score is 0.41 and the level is indicated as green. [10] The projection score will inform us about how well the 3-dimensional PCA plot is representing our data set. More information about projection score can be found in the reference manual.

We can enhance the visualization by selecting the genes that contribute most to the variation over the data set and discarding the genes that only exhibit small (possibly random) fluctuations.

- Move the **Filter by Variance Slider** in the **Statistics Window** and find the setting with the highest projection score (0.43). This can be done by dragging the slider with the mouse.



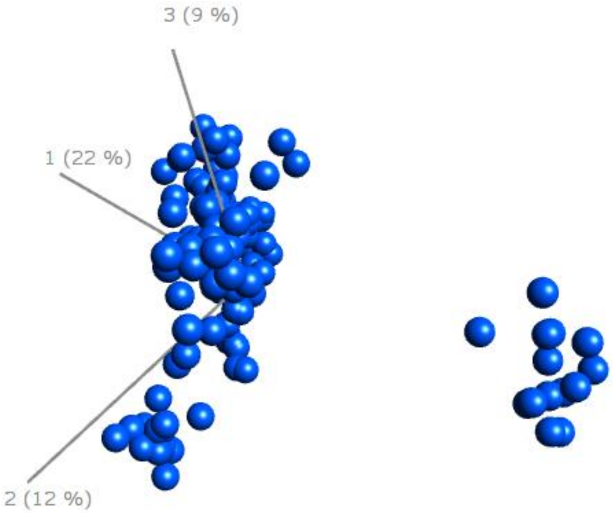
Input: All active variables 22282/22282 vars

Filter by Variance (σ/σ_{max})

0.395 0.43

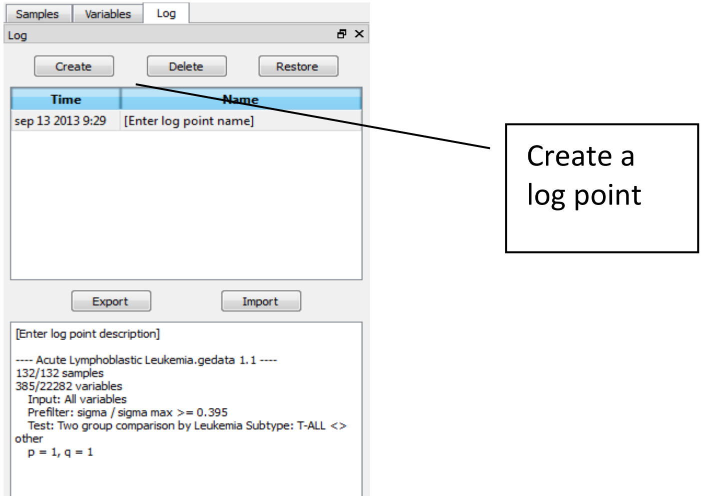
[10] The following rule of thumb is used to color the projection score, above 0.40 is green and below .30 is red, within that interval it is orange/yellow.

Only the genes having a standard deviation of more than (or equal to) 39.5% of the variance of the gene having the largest standard deviation over the samples now take part in the analysis. This happens to be precisely 385 variables (genes) which can be seen in the Status Bar, where it is indicated that only 385 out of 22,282 variables (genes) at the moment participate in the analysis. Clear patterns are now visible in the Plot Window.



By using PCA one makes sure that patients that have resembling gene expression profiles fall close to each other in the plot.

As a parenthesis, remember that you at any time during an analysis of a data set can use the **Log window** to create Log points. By selecting a previously created Log point you can return to that point in the analysis. You can also export (and import) Log points so that you at a later time can return to and retrace the whole analysis.



Time	Name
sep 13 2013 9:29	[Enter log point name]

```
[Enter log point description]

---- Acute Lymphoblastic Leukemia.gedata 1.1 ----
132/132 samples
385/22282 variables
   Input: All variables
   Prefilter: sigma / sigma max >= 0.395
   Test: Two group comparison by Leukemia Subtype: T-ALL <>
other
   p = 1, q = 1
```

Create a log point

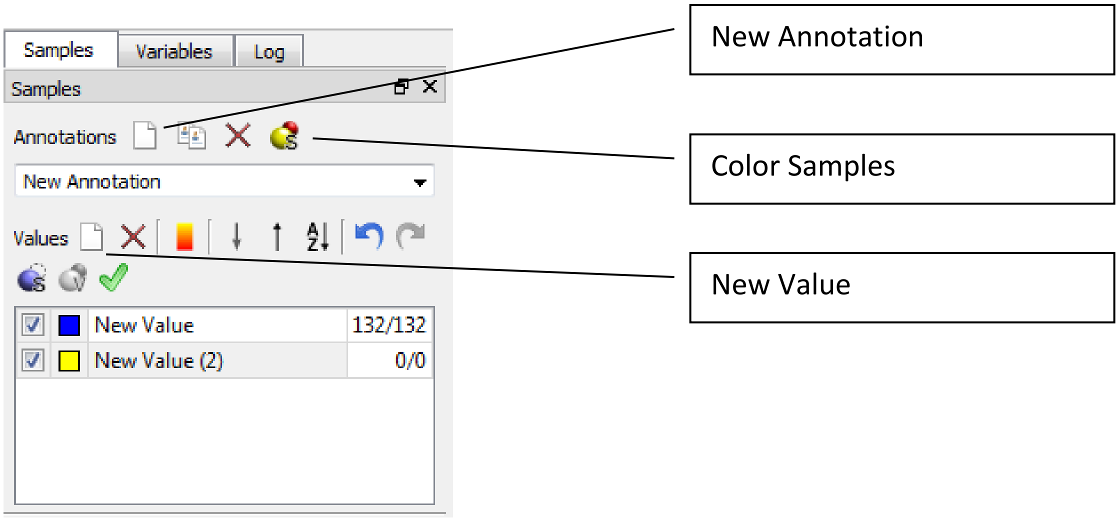
We now return to the ongoing analysis.

As can be seen in the sample PCA plot, the first principal component contains 22 %[11] of the total variance and clearly distinguishes the new group from the rest of the subtypes. In order to more clearly discern structure when plotting the other subtypes we shall now remove the new group from the PCA-analysis.

First we need to create a new sample annotation with several values.

- Select the **New Annotation** button 🗋 in the **Sample Value** panel in the **Sample dock window**.
- Select the **New Value** button 🗋 twice.

It should now look like this.

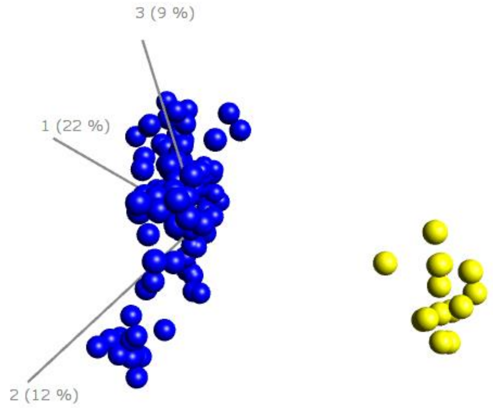


☑	🟦	New Value	132/132
☑	🟨	New Value (2)	0/0

- Color the samples according to the new Annotation by pressing the Quick color button, ●.
- Make sure that the “New Value (2)” group is selected in the **Sample panel**
- Select the mouse tool **Annotate** and circle the new group clockwise

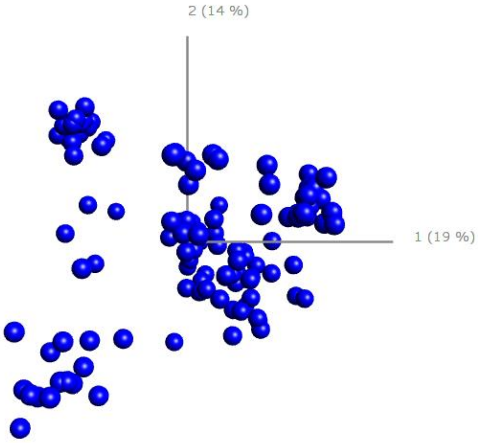
After the circle is closed it will look as below.

[11] The value within parenthesis of the first principal component



- **Uncheck** the checkbox corresponding to New Value 2 in the **Sample Value table** in the **Sample Dock Window**

The PCA is immediately recalculated and we now (after a possible rotation or having the mouse tool in Move and doing clear) have the following plot. Also note that the projection score is recalculated since the number of samples is changed.

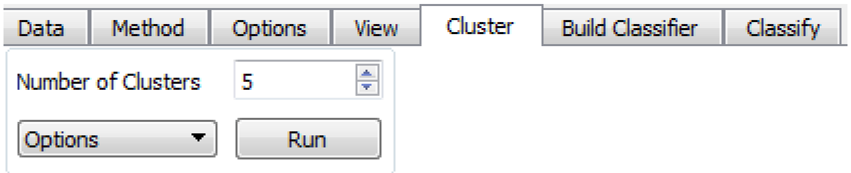


Note that we now have 118 out of 132 samples present in the plot. The text 118/132 Samples is displayed in the **Status Bar** along with the text 288/22282 Variables.

Several subgroups are now clearly discernible and we can continue with the process of creating a new Annotation value, annotate the identified group of samples, remove them from the plot and potentially adjust the projection score.

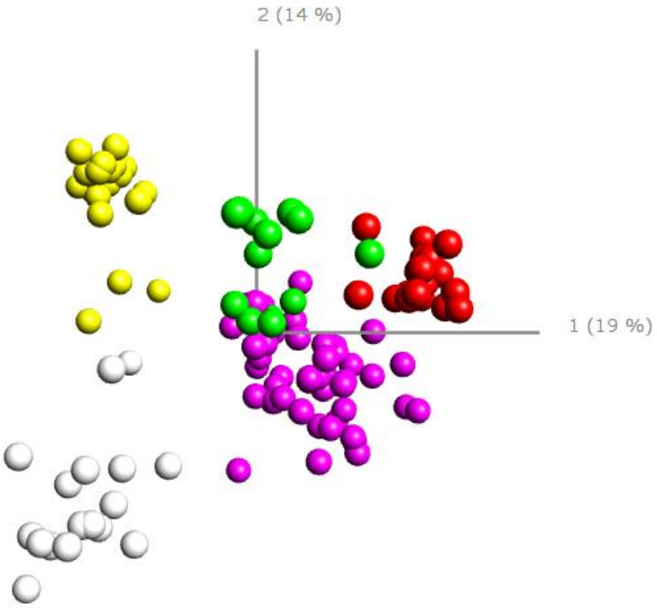
During this exploration we have used PCA combined with variance filtering to identify potential subgroups. An alternative approach would be to use the built in clustering functionality. From the PCA plot we see 5 to 6 potential groups. Let’s apply clustering and look for 5 groups, to see what that unsupervised method would give.

The clustering is controlled from the Cluster tab. Enter 5 and press Run.



Number of Clusters 5

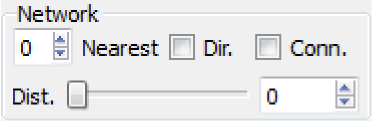
The result is presented as a new Sample annotation. It will be called “k-means 5” after the method and the selected number of clusters. Look in the **Sample Panel**.
After the PCA plot has been colored according to this new Sample annotation (S) it will look as below.



Five potential groups are identified. The silhouette plot is used to assess the quality of the clustering. Positive silhouette values indicates that a sample is close to the other samples in the same group. The silhouette plot is a special configuration of the Bar plot, more about this in the reference manual.

Using Networks and Graphs

You can easily create graphs in QOE connecting samples or variables. When creating graphs the distances involved are always the Euclidean distances in the full space of all active samples or variables. The graphs give you an opportunity to, in a sense, *look* *into higher dimensions*.[12] Using the **Network** textbox in the **Method tab** you can create graphs/networks in many different ways.



Network
0 Nearest □ Dir. □ Conn.
Dist. 0

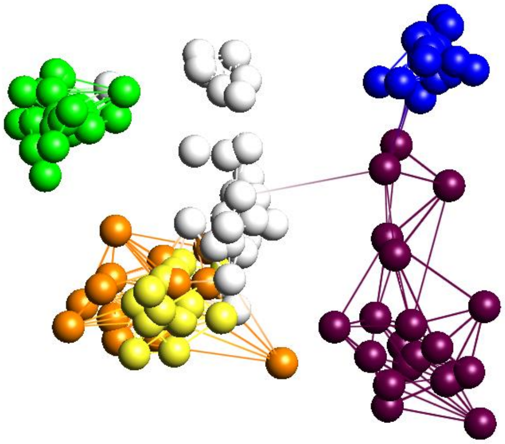
- Uncheck the **Axes** check box in the **Plot Settings** textbox under the **View tab** to see the graph you will create more clearly.

In the **Network** text-box under the **Method tab y**ou select the number of **nearest neighbors** that are, for each distinct sample, to be joined by a graph.

By dragging the slider you create a graph by selecting all neighbors within the **distance** that is selected. Try it.

- Put the **distance slider** to 0 and change the value to 5. You can do this either by using the selection buttons or by writing directly in the textbox and then press Return.

Note again that when one selects the nearest neighbors to create the graphs the distances are computed in the full space of all active variables taking part in the analysis (i.e. in 288 dimensions in the case at hand). After a possible rotation we have the following plot.



An interesting fact manifests itself in the projection above. In the group of diagnostic subtype called “Other”**,** two different subgroups are clearly discernible. The two subgroups do not share any of their 5 closest neighbors. These samples should maybe not be identically classified. We shall now reclassify one of the groups.

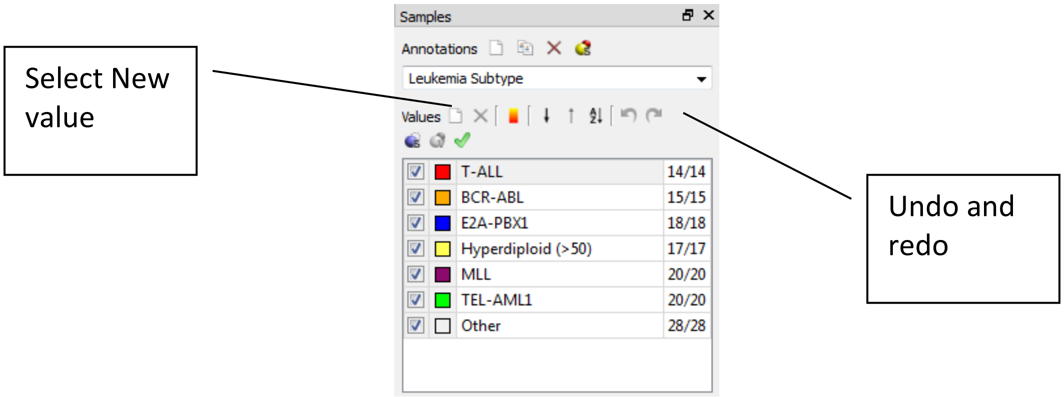
[12] Graphs can also be used in when you work with the multidimensional scaling algorithm Isomap (see the Reference Manual) in OE.

Modifying annotations

It is sometimes convenient to modify the data set you work with, for instance by reclassifying samples, in order to go on and find interesting information. We shall, to begin with, reclassify some of the samples and thereby split one of the diagnostic subgroups into two different subgroups. We do this in order to later be able to find interesting information on for instance which variables (genes) that best discriminate between the two new subgroups.

- Set the network in the **Network** textbox 0 again to clear the graph.
- Select the **New Value** button 🗋 in the **Sample Value** panel in the **Sample dock window**.

A New Value appears in the Value Table and is automatically selected



Select New value

Undo and redo

Samples

Annotations

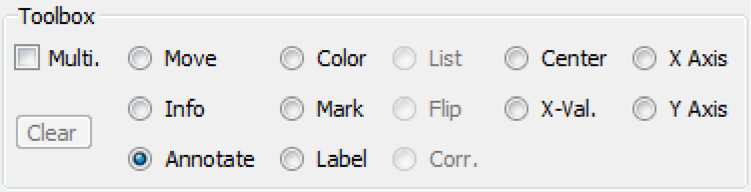
Leukemia Subtype

Values

☑	🟥	T-ALL	14/14
☑	🟧	BCR-ABL	15/15
☑	🟦	E2A-PBX1	18/18
☑	🟨	Hyperdiploid (>50)	17/17
☑	🟪	MLL	20/20
☑	🟩	TEL-AML1	20/20
☑	□	Other	28/28

Note that you can change the name New Value *that appears in the Value Table by simply double clicking in the corresponding text box and entering the preferred name.*

- Select **Annotate** in the **Tool Box** window. This will change the behavior of the Mouse tool to assign samples to a sample annotation value.



Toolbox

□ Multi. ○ Move ○ Color ○ List ○ Center ○ X Axis

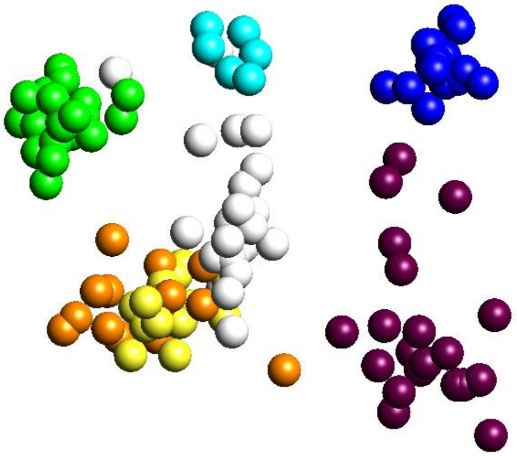
○ Info ○ Mark ○ Flip ○ X-Val. ○ Y Axis

Clear ● Annotate ○ Label ○ Corr.

- Click the samples (the clearly discernible subgroup of **"**Other**"** that is closest to the green subgroup "TEL-AML1"**)** one by one. The samples are then moved to the sample groups "New Value"**.**

*Note that if you happen to move some samples unintentionally you can **undo** your last command by selecting the **Undo** button* ↶ *in the Value Toolbar.*

The plot below is from the middle of the reclassification process. Some of the samples are now annotated in a new group (= light blue) and some are still white.



By selecting **Move** in the **Tool Box**, you can rotate the plot and check that you have marked all samples. If not, you select **Annotate** again and complete the operation. When **Annotate** is selected in the **Tool Box**, you can select **Multi** and you then have the option to select multiple samples by drawing a closed curve around them. You do this by holding down the left mouse button while at the same time moving the pointer (mouse) clock wise around the selected samples to create a closed curve.

Note *The number of variables taking part in the analysis changes when you update the annotation subgroups. This is due to the fact that the **p-value*** is set to 1e-7 *and the set of active variables corresponding to this p-value under the statistical test chosen depends on the subgroups that we choose to discern. With the new subgroup we have added a priori information and it is expected that the ANOVA analysis is influenced.*

- Select **Move** in the **Tool Box**

We shall now open another Plot Window in the Work Space.

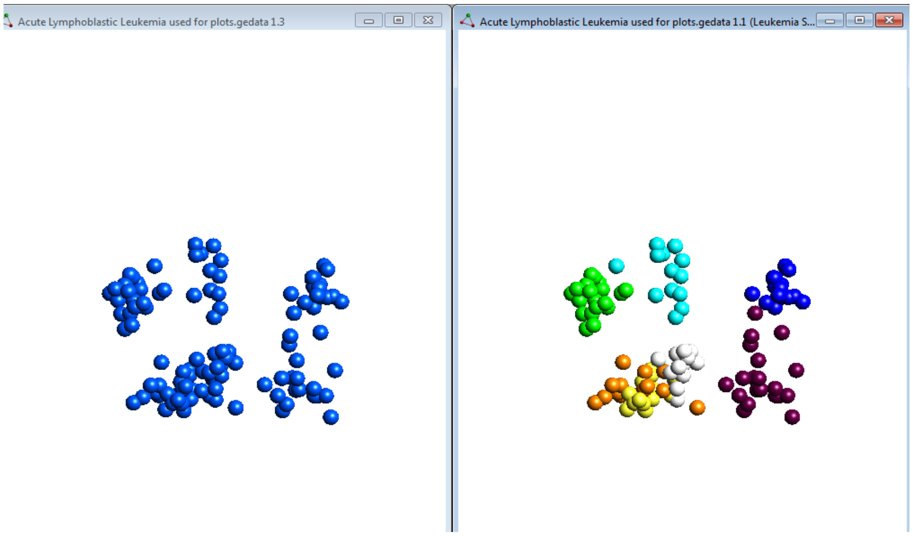
Multiple Plot Windows

You can at any point during the analysis open a new **Plot Window** in the **Work Space**. These new Plot Windows you open can be **synchronized** with the active (highlighted) Plot Window or not. If the Plot Windows are synchronized, they will always share the same active samples and/or variables, but they can, for instance, be colored according to different annotations. This working model is very useful since you can again view several aspects of your data in the same workspace. You **activate** (select) a Plot Window by clicking anywhere in it. The frame of the currently active Plot Window is always highlighted.

- Select **Window** > **New Synchronized Plot** in the **Menu Bar.**

Note that you now have two different Plot Windows open in Qlucore Omics Explorer. You can find them listed under **Window** in the **Menu bar.** You can select which window to display or you can display all windows by choosing **Window** > **Tile** in the **Menu bar**.

- Select **Window** > **Tile** in the **Menu Bar.** We get the following plot.


Acute Lymphoblastic Leukemia used for plots.gedata 1.3
Acute Lymphoblastic Leukemia used for plots.gedata 1.1 (Leukemia S...

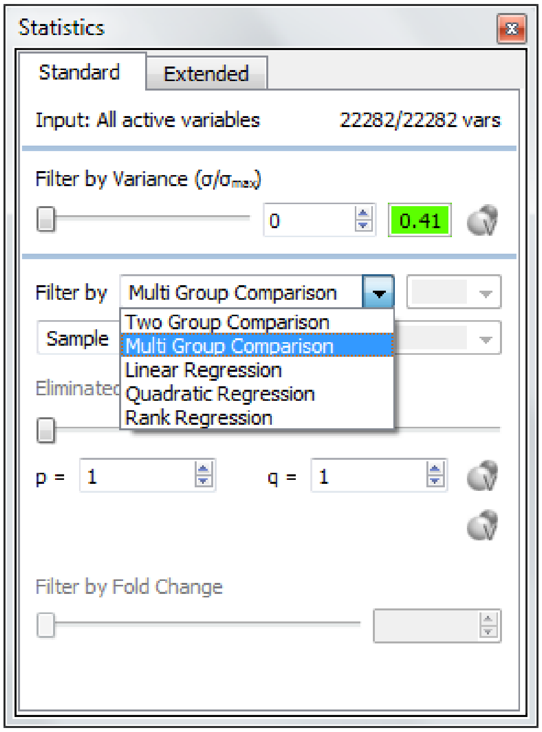
- Make sure that the new Plot Window is active and select **Novel Group** in the **Sample Annotation textbox** in the Sample Dock Window.
- Select the Sample Color Button in the **Sample Annotations Toolbar** to color the samples in the active window according to the Novel group attribute.

We can now see that our **New Value subgroup** in the right window corresponds to the **"Novel group"** in the left window.

Statistical Analysis using ANOVA

To try to separate the different subgroups we shall now select the genes that are most responsible for differentiating between the, a priori, defined diagnostic subtypes.[13] The subtypes are defined by the sample annotation “Leukemia subtype”

- To get a fresh start let’s begin with closing the Leukemia data set and opening it again from the **Help** menu.
- Select the sample annotation “Leukemia subtype” in the sample panel
- Color the samples according to “Leukemia subtype” by using the quick color button, .
- Select the **Multi Group Comparison** in the **Statistics Window**



Input: All active variables 22282/22282 vars

Filter by Variance (σ/σ_{max}) 0 0.41

Filter by Multi Group Comparison

p = 1 q = 1

- Select **Leukemia Subtype** in the corresponding Combo box
- Adjust the p-value slider to a p-value of 1e-30 (or simply type 1e-30 in the textbox and press return).

The corresponding q-value (1.1076e-28) (which can be interpreted as a false discovery rate) is displayed and the 115 genes that now are left in the analysis have a very high

[13] If your objective is to find variables that discriminate between different groups you use the ANOVA functionality in the Statistics Window without first doing variance filtering. This approach is covered in section “Use visual interaction and perform a statistical test” on page 6.

statistical significance. The statistical method used to select the active variables is **ANOVA** and the amount of variation explained by diagnostic subtype is called the R2-value.

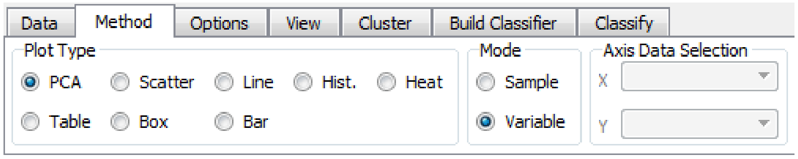
The statistical significance can be verified in QOE by using ***randomization*** and ***permutation*** of your data set. This functionality can be found in the **Data tab**.

In the Status Bar you can now see that 115 out of the total 22,282 variables are active, i.e. participate in the analysis. What we see in the current PCA plot is a three-dimensional projection from 115 dimensions.

Working with Variables

Although we have, strictly speaking, been working with the variables all the time, since we have filtered the data, we shall now have a look at them explicitly.

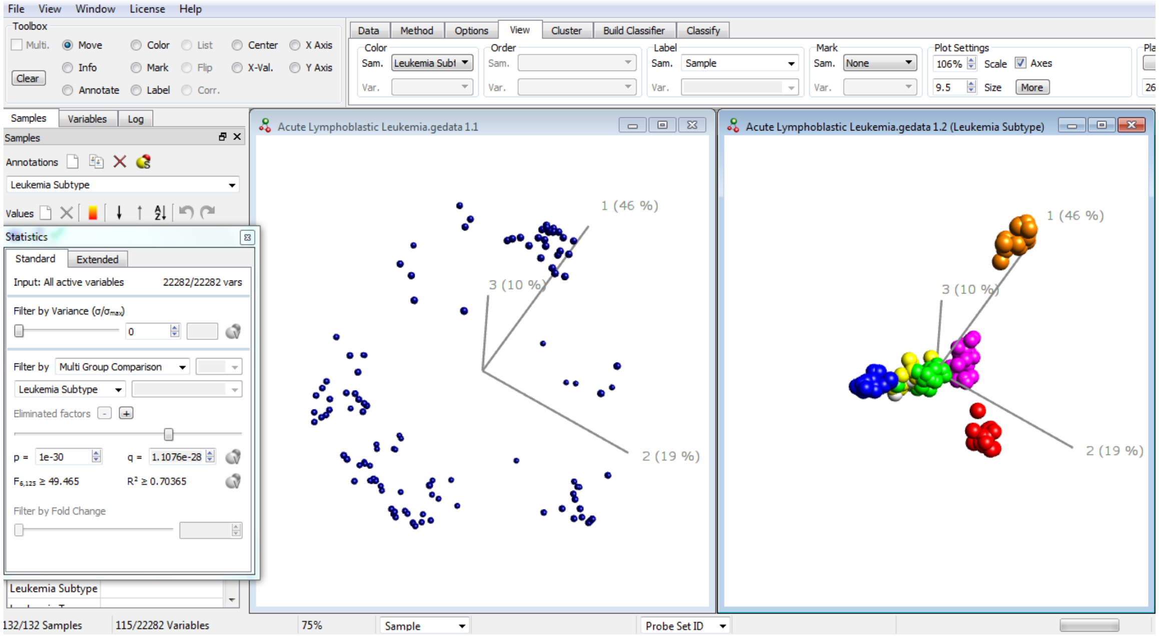
- Create a new Synchronized Plot Window by selecting the **Window** menu and **New Synchronized Plot**.
- Select **Tile** in the **Window** menu to see both plots
- Make sure that the Plot Window to the left is active and then select Plot Type **PCA** and Mode **Variable** in the **Method tab** in order to display a variable PCA plot.



Plot Type: ● PCA ○ Scatter ○ Line ○ Hist. ○ Heat ○ Table ○ Box ○ Bar

Mode: ○ Sample ● Variable

This gives the following two plots



Color Sam. Leukemia Subt
Order Sam.
Label Sam. Sample
Mark Sam. None
Plot Settings 106% Scale ☑ Axes 9.5 Size More

Samples
Leukemia Subtype

Input: All active variables 22282/22282 vars
Filter by Variance (σ/σ_{max}) 0
Filter by Multi Group Comparison
Leukemia Subtype
p = 1e-30 q = 1.1076e-28
$F_{6,125} \geq 49.465$ $R^2 \geq 0.70365$

Leukemia Subtype

Acute Lymphoblastic Leukemia.gedata 1.1

Acute Lymphoblastic Leukemia.gedata 1.2 (Leukemia Subtype)

132/132 Samples 115/22282 Variables 75% Sample Probe Set ID

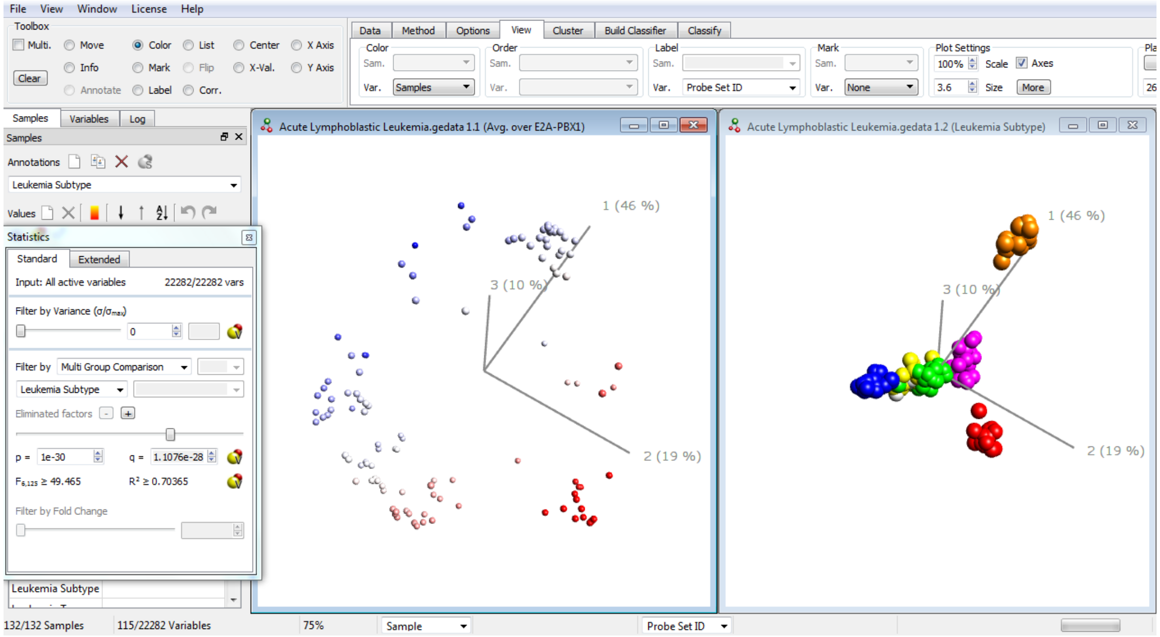
In the left Plot Window above you see a PCA plot of the 115 active variables participating in the analysis at the moment.

- Select the left window by clicking anywhere in it, to activate it.
- Select **Color Var.** under the **View Tab**.
- Select to color the variables “by data for one or more samples” and then select the Leukemia Subtype annotation and finally the “E2A-PBX1” group.

There are a number of ways to color variables (Solid, variance, R^2 ,…) see the **Variable Color options** list for all options.

The variables are now colored according to the mean expression level in the subtype group selected in the Value Table for the Samples (“E2A-PBX1”). Red means highly expressed, i.e. they are up-regulated in the chosen Sample group and green corresponds to down-regulated genes.

In the left Plot Window below you see a three-dimensional PCA plot of the 115 variables, taking part in the analysis at the moment, colored according to their mean expression level in the Sample subtype group “E2A-PBX1”. Notice that since we have chosen to work with synchronized plots, the highly expressed variables in the “E2A-PBX1” group are found in the same direction in the plot as the “E2A-PBX1” group itself.


File View Window License Help
Toolbox
Data Method Options View Cluster Build Classifier Classify
Acute Lymphoblastic Leukemia.gedata 1.1 (Avg. over E2A-PBX1)
Acute Lymphoblastic Leukemia.gedata 1.2 (Leukemia Subtype)
1 (46 %)
2 (19 %)
3 (10 %)

Statistics: Standard

Input: All active variables 22282/22282 vars

Filter by Variance (σ/σ_{max}) 0

Filter by Multi Group Comparison, Leukemia Subtype

p = 1e-30   q = 1.1076e-28

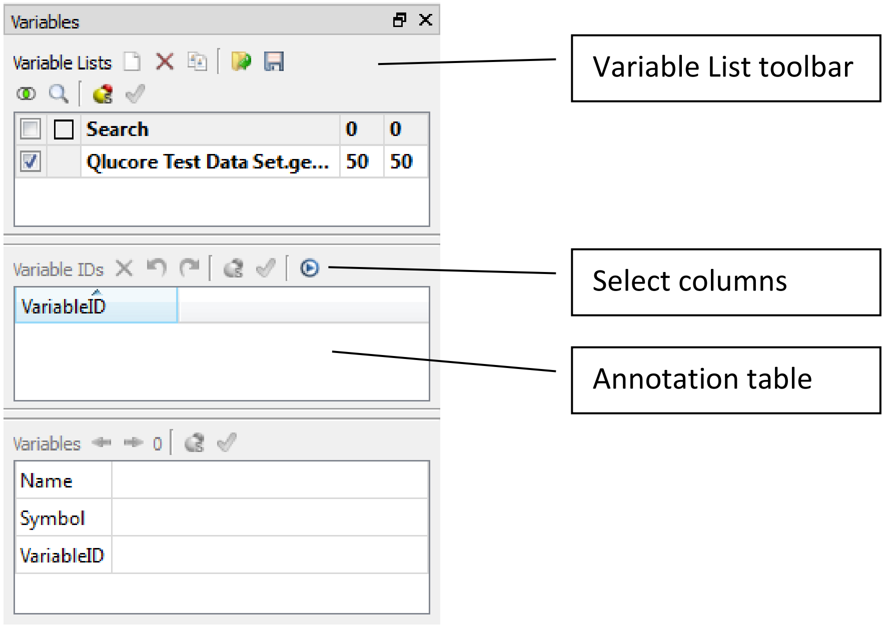
$F_{6,125} \geq 49.465$ $R^2 \geq 0.70365$

132/132 Samples   115/22282 Variables   75%   Sample   Probe Set ID

- Select **View>Dock Windows>Color Legend** to see the color scale

We shall now create a list of the genes that are most up-regulated in the “E2A-PBX1” group.

- Select the **Variables Dock Window**



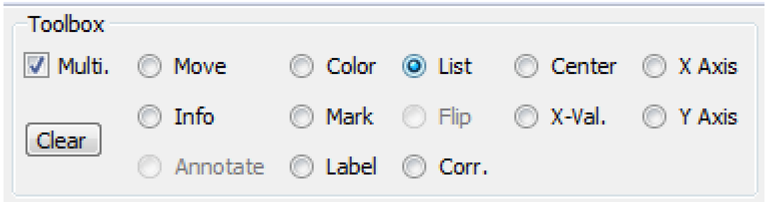
☐	☐	Search	0	0
☑		Qlucore Test Data Set.ge...	50	50

VariableID

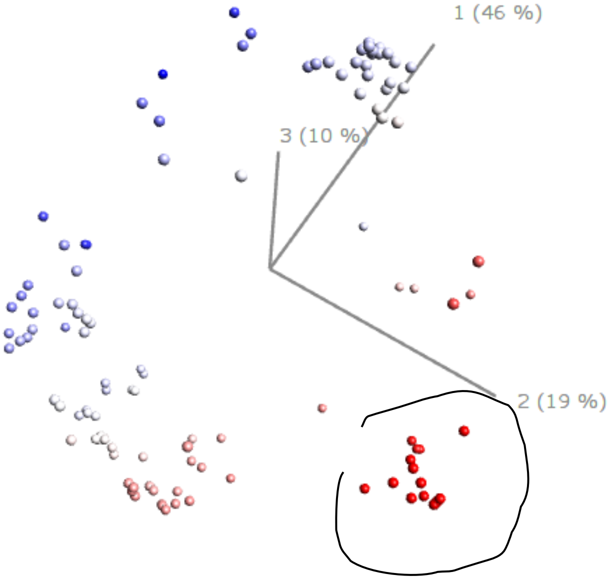
Name	
Symbol	
VariableID	

- Select the left window (**Variable PCA**) by clicking anywhere in the window.

- Click the **New** button in the **Variable Lists toolbar** to create a new variable list
- Select **List** in the **Tool Box** (in order to be able to make this selection, the right plot window must be active).



Draw a closed curve clock wise around some of the genes that are most expressed for the group “E2A-PBX1”. You do this by holding down the left mouse button while at the same time moving the mouse tool tip clock wise around the selected genes to create a closed curve.



The new **variable list** is available in the **Variable List Table,** displaying the selected genes.

Note that the annotations for the selected genes in the variable name list are found in the **Variable Annotation Table.**

The **List tool** works in plot types including variables. This can be a variable PCA plot, a heatmap or a scatter plot.

With the **Select Columns button** in the variable dock window you can select what information you want to present in the Variable Table. You can for instance get the p- and q-values for every individual variable.

Create a copy of the list you have created using the button and name the list to for instance “Separating E2A-PBX a selection”, finally select the **Save** button to save the variable list for use at a later occasion.

With the **Open** button , you can import an already stored variable list. It is possible to have many variable lists open at the same time. The lists can either be created within QOE, as we just have done, or be a list of variable identifiers created from other sources (gene ontology categories, pathways,..).

Building a classifier

An alternative way, to statistical tests, to identify variables (features) of key interest for instance when the objective is to do biomarker discovery, is to create a classifier and observe which variables that are selected. This list of variables is a good starting point to understand which variables that are the best potential biomarkers.

Building a classifier is done from the **Build Classifier tab**. The output is the classifier, an extensive report and a variable list of the selected variables.

Read more about the classifier functionality in specific documentation found on www.qlucore.com/documentation.aspx

Utilizing System biology information (GSEA and GO)

Pathway analysis, or gene set analysis, is a collective name for methods aimed at statistical analysis of a collection of genes, rather than single genes, in a given data set. Typically, genes are grouped together in a collection (or a gene set) if they have something in common, for example, if they are part of the same biological pathway or if they are all located close to each other along the genome.

To perform pathway (gene set) analysis, two components are needed: data set, and one or several predefined gene sets (that is, the gene sets should not be defined based on the values in the data set). Gene set definitions are often acquired from open online repositories such as mSigDB and Reactome, or from commercial products specialized in providing manually curated pathway information.

Gene Set Enrichment Analysis (GSEA)

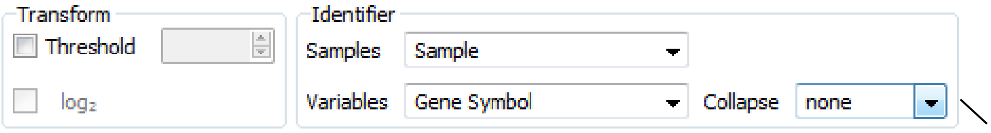
We have so far only used the content of the data file (including annotations) to perform the analysis of the experiment data. Qlucore also offers the GSEA workbench as a tool to analyze the outcome of a statistical test in the context of other lists (gene sets). Read more about GSEA as a method in *Subramanian, Tamayo, et al. 2005 Proc Natl Acad Sci U S A 102(43):15545-50.*

To demonstrate the GSEA Workbench and the GO-browser we will re-start the analysis.

- Close all open data sets in QOE.
- Open the Acute Lymphoblastic Leukemia data set from **Help > Example files**

The example gene sets provided in QOE are built around Gene Symbols as the unique identifier for genes whereas the Leukemia data set we have used so far is based on probeset IDs (Affymetrix) as the unique identifier for each variable (gene). A comparison without taking this into account will give zero matches between the two information sources.

Select the **Data** tab and the **Identifier** box. Change **Variable Identifier** to "Gene Symbol". Watch how the two columns, for the Active list, in the variable list panel are updated to reflect the identifier change. The first column is the number of unique elements (Gene Symbols) in the data set and the second column is the number of matches. Since we have not done any filtering or selections the number of matches equals the number of probes sets in the data set (22,282).



Transform
□ Threshold
□ $\log_2$

Identifier
Samples: Sample
Variables: Gene Symbol
Collapse: none

Collapse

To get unique variables for each Gene Symbol we need to collapse the variables which have one or more probe sets attached. Select to collapse based on **Average** in the drop down box next to the selection of Variable identifier. The updated data set now include 13,262 variables that each of them matches a Gene Symbol.

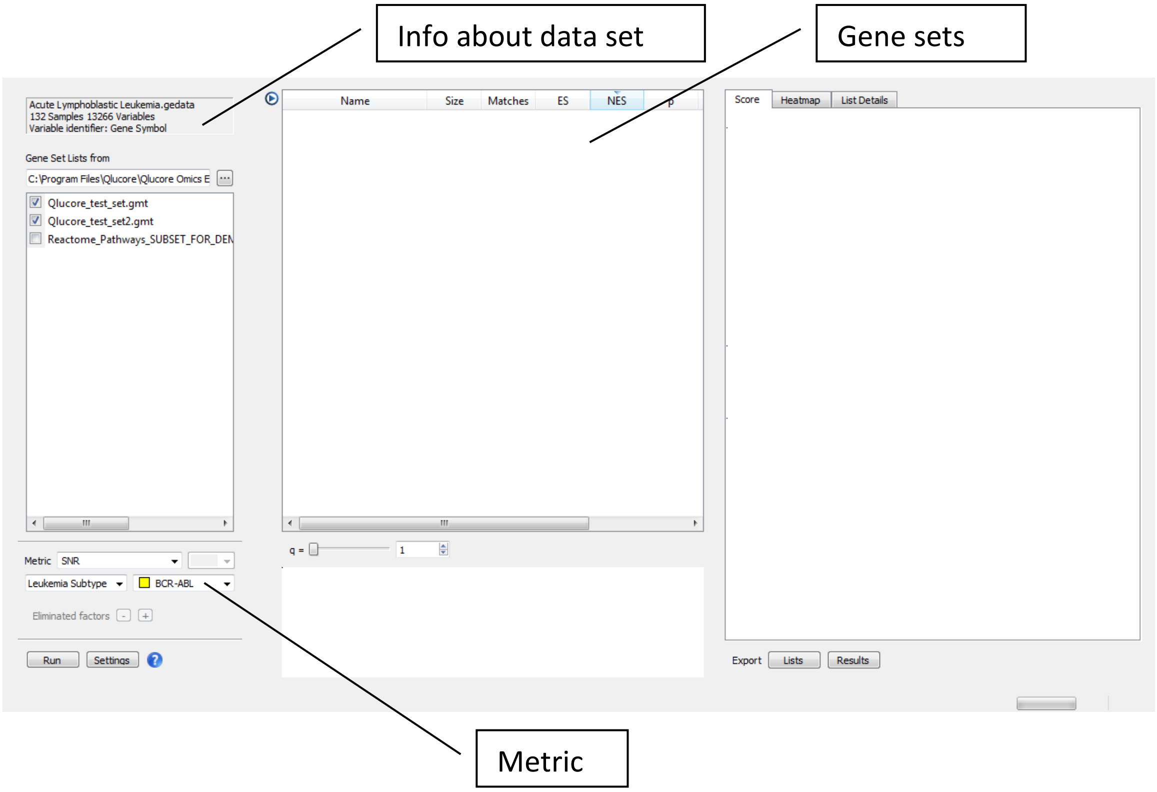
Start the **GSEA Workbench** (**View menu)**. A new window will open. QOE is shipped with three example gene sets for demonstration purposes. When the GSEA Workbench is started it makes a copy of the data in the active Data set. Relevant statistical settings are also copied from the Statistics dialog.

Note: Since the ***GSEA Workbench*** *works with a copy of the data set you can continue to work and analyze your data set in the QOE Main Window.*

If you use the default settings you will see a screen as below. Now change the settings according to the steps below.

- Make sure both Qlucore Test Gene Sets are selected
- Change the Metric to be SNR on Leukemia Subtype and group E2A-PBX1.
- Press Run to start the calculation of the Enrichment scores.

The Enrichment scores are calculated using the metric selected



Info about data set

Gene sets

Metric

Acute Lymphoblastic Leukemia.gedata
132 Samples 13266 Variables
Variable identifier: Gene Symbol

Gene Set Lists from
C:\Program Files\Qlucore\Qlucore Omics E

☑ Qlucore_test_set.gmt
☑ Qlucore_test_set2.gmt
□ Reactome_Pathways_SUBSET_FOR_DEN

Metric: SNR
Leukemia Subtype: BCR-ABL

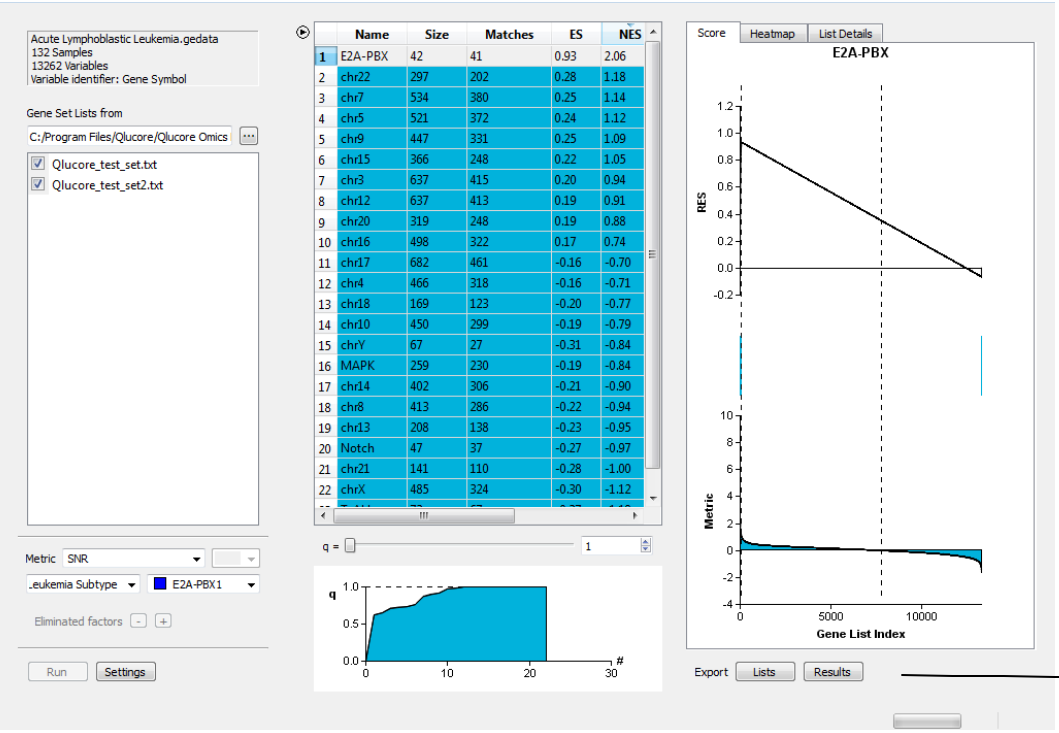
Name	Size	Matches	ES	NES	p

q = 1

The Enrichment Score is calculated for all gene sets and the result list in the middle is ordered according to the Enrichment Score.

Press list nr 1. You will get the results as below.

The first list, called E2A-PBX has the highest Normalized Enrichment score (2.06). The graphs to the right show the results. Since the metric was chosen to rank the data set according to the SNR for E2A-PBX1 it is not a big surprise that we get a very clear plot for the example gene list called E2A-PBX1.



Acute Lymphoblastic Leukemia.gedata
132 Samples
13262 Variables
Variable identifier: Gene Symbol

Gene Set Lists from: C:/Program Files/Qlucore/Qlucore Omics

☑ Qlucore_test_set.txt
☑ Qlucore_test_set2.txt

Metric: SNR
Leukemia Subtype: E2A-PBX1

	Name	Size	Matches	ES	NES
1	E2A-PBX	42	41	0.93	2.06
2	chr22	297	202	0.28	1.18
3	chr7	534	380	0.25	1.14
4	chr5	521	372	0.24	1.12
5	chr9	447	331	0.25	1.09
6	chr15	366	248	0.22	1.05
7	chr3	637	415	0.20	0.94
8	chr12	637	413	0.19	0.91
9	chr20	319	248	0.19	0.88
10	chr16	498	322	0.17	0.74
11	chr17	682	461	-0.16	-0.70
12	chr4	466	318	-0.16	-0.71
13	chr18	169	123	-0.20	-0.77
14	chr10	450	299	-0.19	-0.79
15	chrY	67	27	-0.31	-0.84
16	MAPK	259	230	-0.19	-0.84
17	chr14	402	306	-0.21	-0.90
18	chr8	413	286	-0.22	-0.94
19	chr13	208	138	-0.23	-0.95
20	Notch	47	37	-0.27	-0.97
21	chr21	141	110	-0.28	-1.00
22	chrX	485	324	-0.30	-1.12

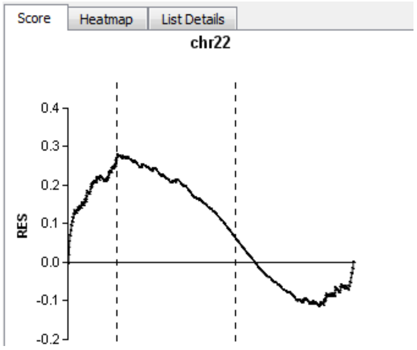
q = 1

Export

A general guide for interpretation is that the data set you are analyzing, based on the chosen metric, is showing the highest degree of commonality with the gene set list with the highest Enrichment score.

Select the second list in the table, i.e. chr22.

This will give a more normal plot showing how the score is growing up to the Enrichment score of 0.28.



There are two Export options:

Lists: Will export the content of the selected gene set lists to QOE. The lists will be visible in the Variable List Table.

Results: Will export all plots and result lists into a folder of your choice.

Close the **GSEA Workbench** to prepare for the next step

GO Browser

To show how system biology information such as a gene ontology can enhance the analysis we will start the **GO Browser**. You can use different ontology and association files as input. By default, a generic GO Slim ontology and a Gene Association file for humans are included. For the latest versions and for other ontologies visit www.geneontology.org[14]. These files are updated continuously and the results in the example below might differ from what you experience performing the steps using newer files.

- Uncheck the "T-ALL" group in the Sample panel.
- In the statistics dialog filter by Multi Group Comparison and Leukemia subtype
- Change to a Variable PCA plot
- Filter to a p-value of 1e-15. This will give you 345 active variables.
- Start the **GO Browser** from **View > GO Browser**[15].

Search for "kinase" in the GO Browser window. We only have one hit, GO:0016301, which is a sub category of the molecular function ontology (GO:0003674).

By selecting a row in the list of search results the content of that category is shown in the upper right window.

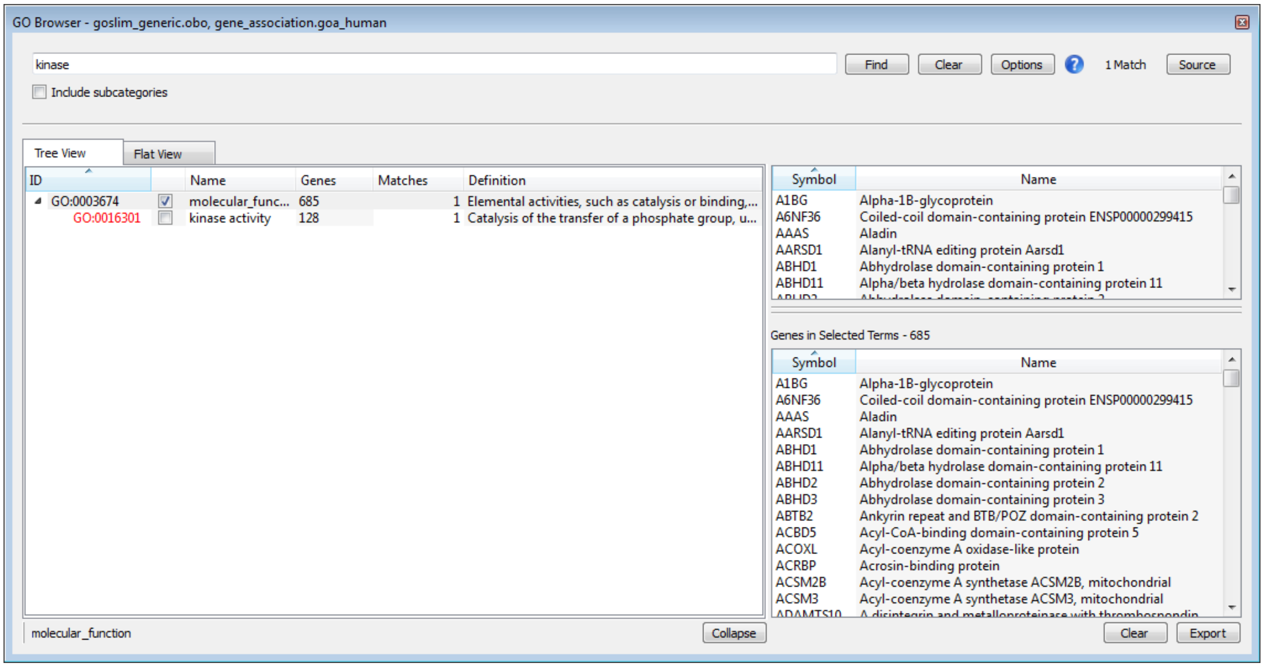
Check the checkbox to the right of the GO term (GO:0016301). You will now see the 180 genes matching the selected term shown in the window to the lower right. The lower right window will show the sum of all selected terms. These genes can be exported to the Qlucore Omics Explorer **Variable List Table** interface.

- Press the Export button to export the list.
- Uncheck the GO:0016301 and check GO:0003674
- Press the Export button again

Switch to QOE main window. You will now see two new lists in the **Variable List Table** named according to the GO search terms.

[14] The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data.

[15] Make sure to have a proper Ontology as input file. For details please see the reference manual.



GO Browser - goslim_generic.obo, gene_association.goa_human

kinase

☐ Include subcategories

ID		Name	Genes	Matches	Definition
GO:0003674	☑	molecular_func...	685	1	Elemental activities, such as catalysis or binding,...
GO:0016301	☐	kinase activity	128	1	Catalysis of the transfer of a phosphate group, u...

Symbol	Name
A1BG	Alpha-1B-glycoprotein
A6NF36	Coiled-coil domain-containing protein ENSP00000299415
AAAS	Aladin
AARSD1	Alanyl-tRNA editing protein Aarsd1
ABHD1	Abhydrolase domain-containing protein 1
ABHD11	Alpha/beta hydrolase domain-containing protein 11

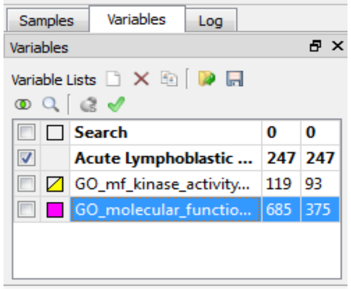
Genes in Selected Terms - 685

Symbol	Name
A1BG	Alpha-1B-glycoprotein
A6NF36	Coiled-coil domain-containing protein ENSP00000299415
AAAS	Aladin
AARSD1	Alanyl-tRNA editing protein Aarsd1
ABHD1	Abhydrolase domain-containing protein 1
ABHD11	Alpha/beta hydrolase domain-containing protein 11
ABHD2	Abhydrolase domain-containing protein 2
ABHD3	Abhydrolase domain-containing protein 3
ABTB2	Ankyrin repeat and BTB/POZ domain-containing protein 2
ACBD5	Acyl-CoA-binding domain-containing protein 5
ACOXL	Acyl-coenzyme A oxidase-like protein
ACRBP	Acrosin-binding protein
ACSM2B	Acyl-coenzyme A synthetase ACSM2B, mitochondrial
ACSM3	Acyl-coenzyme A synthetase ACSM3, mitochondrial
ADAMTS10	A disintegrin and metalloproteinase with thrombospondin

molecular_function

The second column in the Variable List Table enables the option to color variables according to any list(s).

- Press the color box for the first list and then press the color list for the second list.

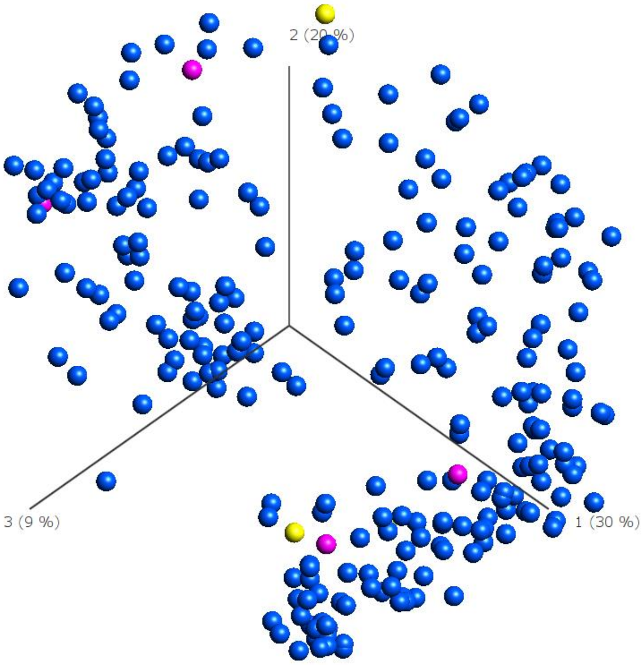


☐	☐	**Search**	**0**	**0**
☑		**Acute Lymphoblastic ...**	**247**	**247**
☐	▨	GO_mf_kinase_activity...	119	93
☐	■	GO_molecular_functio...	685	375

Note: The first GO list includes 119 unique items which matches precisely 93 genes in this specific data set whereas the second list includes 685 items which matches 375 genes.

The plot should look like something as below, where all genes included in the list GO_molecular_function_GO_0003674 are colored purple and the genes in the list GO_mf_kinase_activity_GO_0016301 are colored yellow. If a list gets a color with a diagonal in the color box it means that the lists are overlapping. The list last used for coloring will prevail.

This work model is an excellent way to combine system biology information from different lists with the conclusions from the ongoing study. It is possible to color the variable plots to any number of lists.



Note: The variable coloring works also in heatmaps and scatter plots.

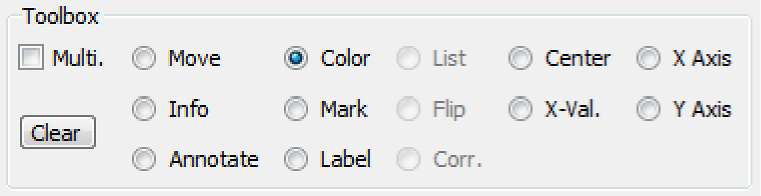
To study only the variables included in the two gene ontology based lists is straightforward.

- In the **Variable List Table** check the first column for the two lists.

The plot will be updated and only include yellow and purple colored genes, there will be 15 active variables.

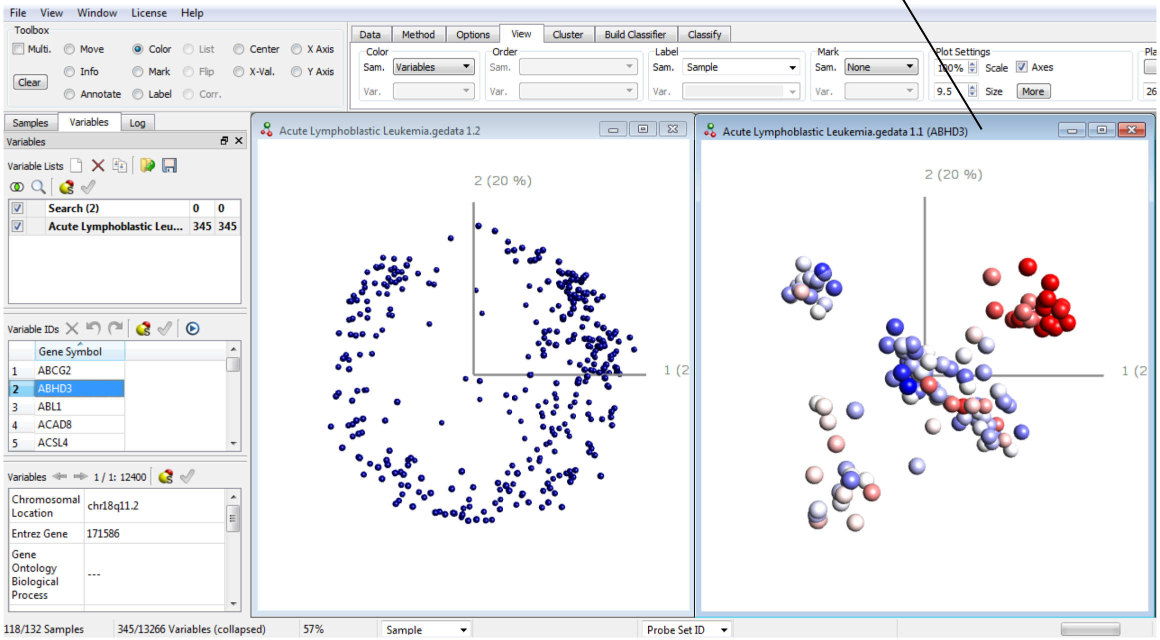
The color tool mouse tool is also powerful to understand more about the data. To illustrate this further open a synchronized sample PCA plot (**Windows > New Synchronized Plot**) and then select to tile (**Windows > Tile**)

- Select **Color** in the **Tool Box** (no multi)**.** Make sure that the Sample PCA plot is active and then select (in the Variable Window) a variable in the variable PCA plot



You now get the **Samples** colored according to the expression level for the selected gene for each sample. In the example below you observe that sample data is colored according to the gene ABHD3. Red means high and blue corresponds to low levels.

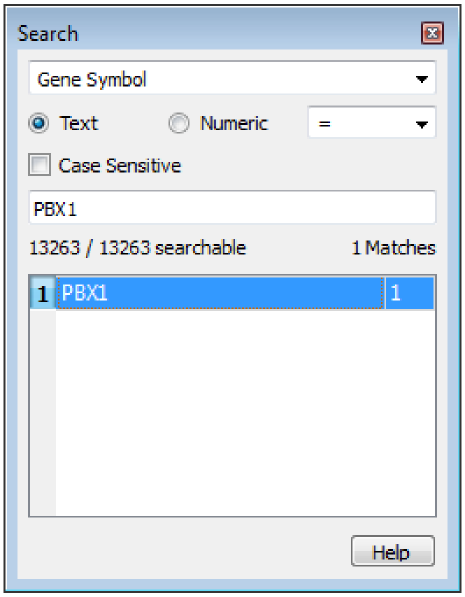
Colored according to ABHD3


File View Window License Help
Toolbox
Data Method Options View Cluster Build Classifier Classify
Color Sam. Variables Var.
Order Sam. Var.
Label Sam. Sample Var.
Mark Sam. None Var.
Plot Settings 100% Scale Axes 9.5 Size More
Samples Variables Log
Variables
Variable Lists
Search (2) 0 0
Acute Lymphoblastic Leu... 345 345
Variable IDs
Gene Symbol
1 ABCG2
2 ABHD3
3 ABL1
4 ACAD8
5 ACSL4
Variables 1 / 1: 12400
Chromosomal Location chr18q11.2
Entrez Gene 171586
Gene Ontology Biological Process ---
Acute Lymphoblastic Leukemia.gedata 1.2
Acute Lymphoblastic Leukemia.gedata 1.1 (ABHD3)
2 (20 %)
1 (2
118/132 Samples
345/13266 Variables (collapsed)
57%
Sample
Probe Set ID

By selecting the different variables (one by one) either in the variable PCA plot or in the variable list one sees the expression level of the selected genes for each patient.

We shall now find genes that are correlated with a specific gene. We select the gene PBX1.

- Use the Search tool to find the PBX1 gene. Select the 🔍 button and enter PBX1 in the search dialog when you search in the "Gene Symbol" annotation.



Gene Symbol

● Text ○ Numeric =

□ Case Sensitive

PBX1

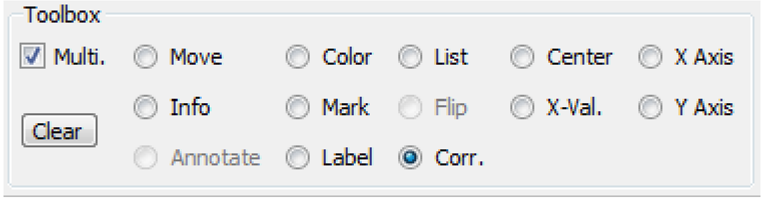
13263 / 13263 searchable 1 Matches

1	PBX1	1

The annotation to search in

1 match is found.

Select **Corr.** in the **Tool Box**



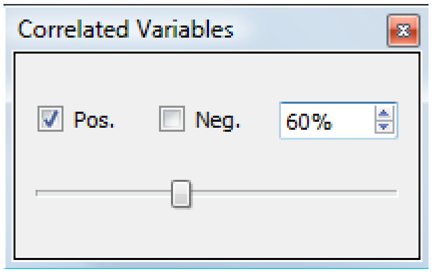
☑ Multi. ○ Move ○ Color ○ List ○ Center ○ X Axis

○ Info ○ Mark ○ Flip ○ X-Val. ○ Y Axis

○ Annotate ○ Label ● Corr.

Press the Search list in the Variable Table View. The PBX1 gene will be added to the variable PCA plot and it will be marked.

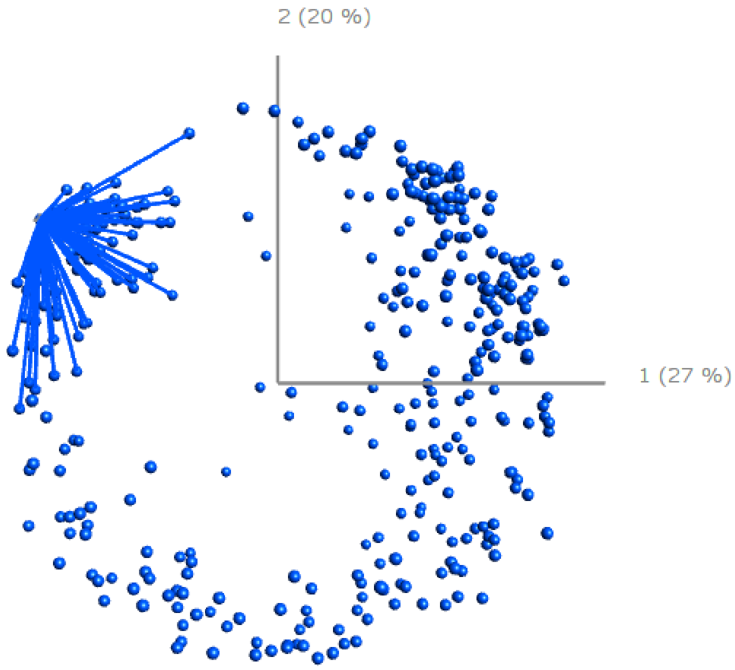
In the **Correlated Variable Box** you can now choose the correlation level and if you want to include both positive and negative correlations.



☑ Pos. □ Neg. 60%

- **Select** 60% and only positive correlations

In the Variable Plot Window you can now see that all variables that have a correlation of more than 60% with PBX1 are connected with a line. All other active variables are also present. To only see the correlated variables move the variance slider to the right.



Further analysis and exploration

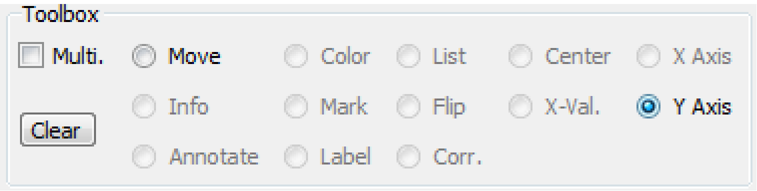
At this point in the analysis we have covered a broad range of functionality and you have familiarized yourself with selection methods, coloring options, synchronized plots and much more. In the following sections we will more briefly highlight additional functionality. We start with the **Box plot.**

Box plot

- Close the **Variable PCA** plot
- Close the **Correlated Variable Box**
- Check the active Variable list to include all variables
- Change to Filter by Two Group comparison and E2A-PBX1
- Select the **Method tab** and change the plot type to **Box**.

To populate the plot, data for the X-axis and Y-axis need to be selected. Navigate to the X Axis drop down box in the **Axis Data Box** in the **Method Tab** and select the Sample Annotation “Leukemia Subtype”.

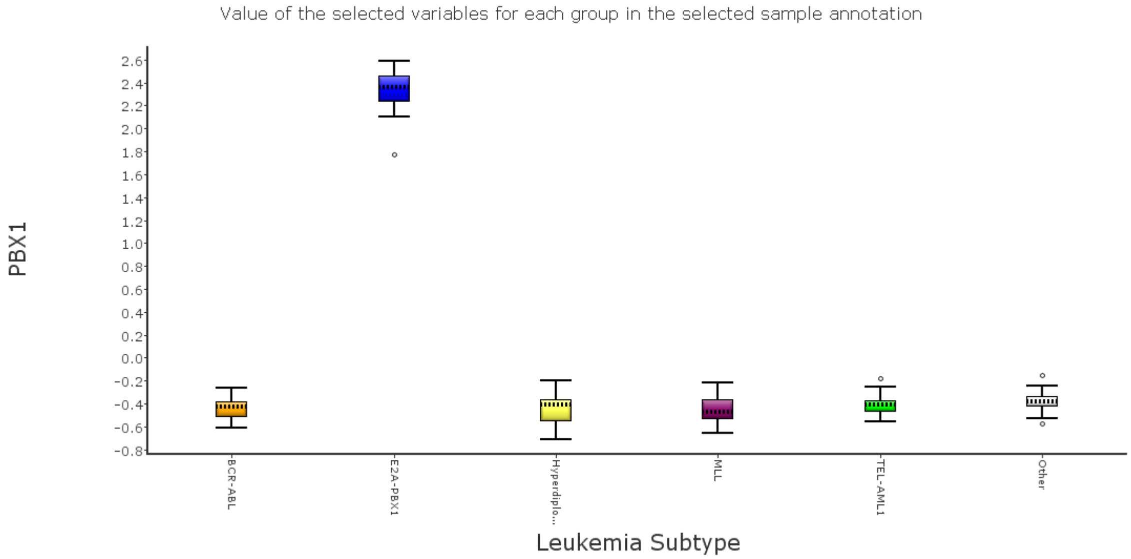
- Select the **Y Axis** tool in the **Tool Box**



- Select the Variable in **Search list** in the **Variable Window**.

You should get a plot as below. As expected the values for this variable are high in the Blue group (E2A-PBX1) since that was how we selected the variable from the start. Each box is calculated based on the samples in the respective sub group. The parts of the box are defined according to the following

- The dotted line is the Mean value
- The upper limit of the box is 75 percentile (default)
- The lower limit of the box is the 25 percentile (default)
- By default the box whiskers are set at the lowest data point value still within 1.5 times the box range of the lower box limit, and at the highest data point value still within 1.5 times the box range of the upper box limit. The circle represent a potential outlier and they are defined by data elements outside the whiskers.



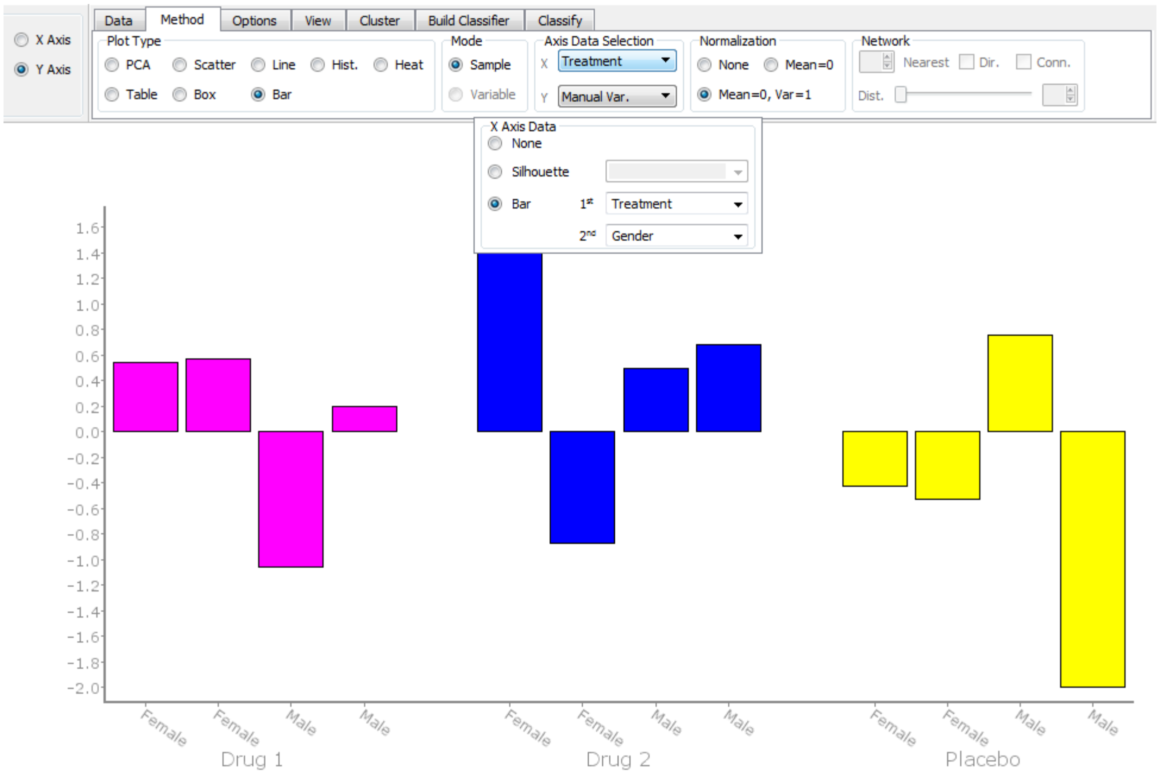
In the **Options Tab** the **Box Limits** can be changed and adjusted to your needs.

Bar plot

The **Bar plot** can be configured in several ways and it also supports operations on the groups (such as average)

The **Bar plot** is controlled from the **Method tab** where X-axis and Y-axis content is selected. Data operations are controlled from the **Options tab.** From the **View tab** are the visual configurations operated.

The first plot below shows the Qlucore Test data set. The X-axis is first ordered according to the annotation “Treatment” which has three values (“Drug1”, “Drug2” and “Placebo”), the second order is by the annotation “Gender”. This means that within each “Treatment” subgroup the bars are ordered according to “Gender”, from the plot we see that “Female” samples are first and then “Male”.



○ X Axis
● Y Axis

Plot Type: ○ PCA ○ Scatter ○ Line ○ Hist. ○ Heat ○ Table ○ Box ● Bar

Mode: ● Sample ○ Variable

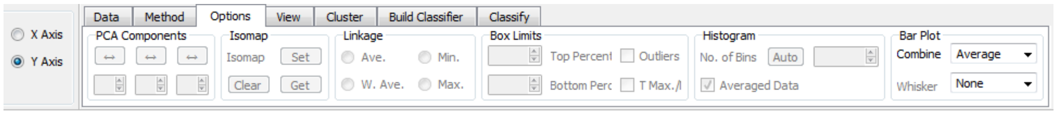
Axis Data Selection: X Treatment; Y Manual Var.

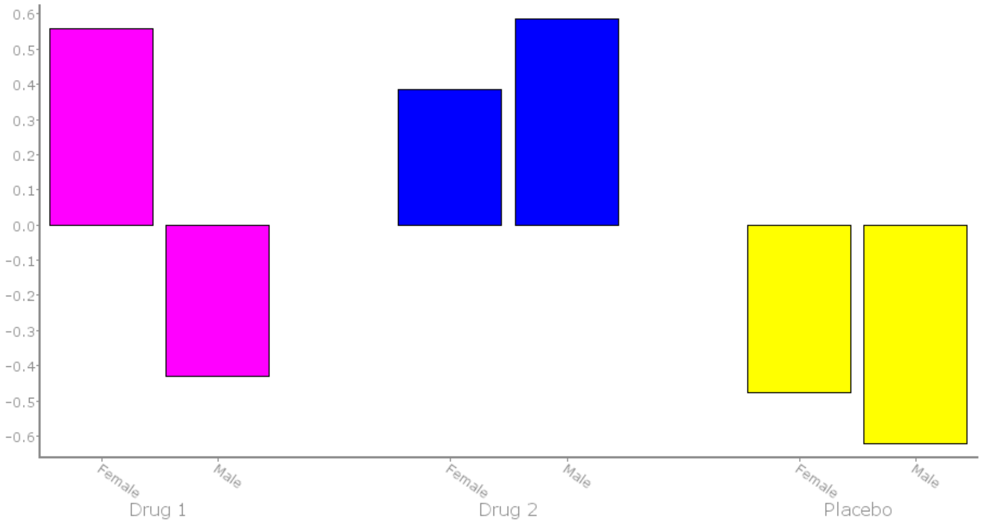
Normalization: ○ None ○ Mean=0 ● Mean=0, Var=1

Network: Nearest □ Dir. □ Conn. Dist.

X Axis Data
○ None
○ Silhouette
● Bar 1st Treatment
2nd Gender

In the **Options tab** data different Combine operations on the data can be defined. In the plot below data is averaged. The operation always applies to the second annotation selected in the **Method tab**.



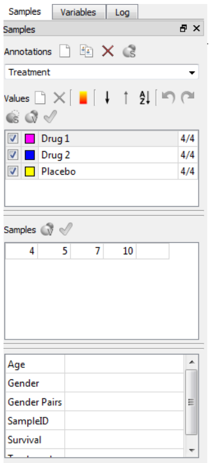


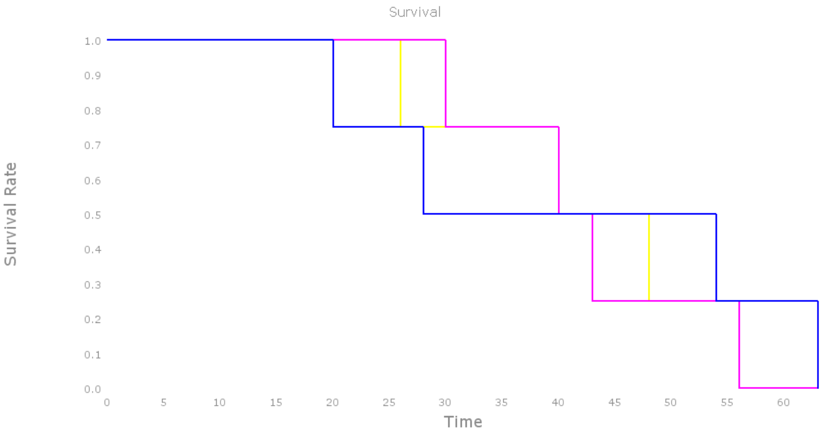
Line plots and Kaplan Meier survival plot

To generate a **Kaplan Meier plot** a Sample annotation containing survival time is required. Censoring information can also be used. This should then be available as a second Sample annotation.

Select the **Line plot** and the **Kaplan-Meier** option in the X-Axis Data Selection in the **Method tab**. Also define if the data shall be organized into different groups.

Below is an example from the Qlucore test data set. Survival for the patients in the three different “Treatment” groups (“Drug1”, “Drug2” and “Placebo”) is presented.



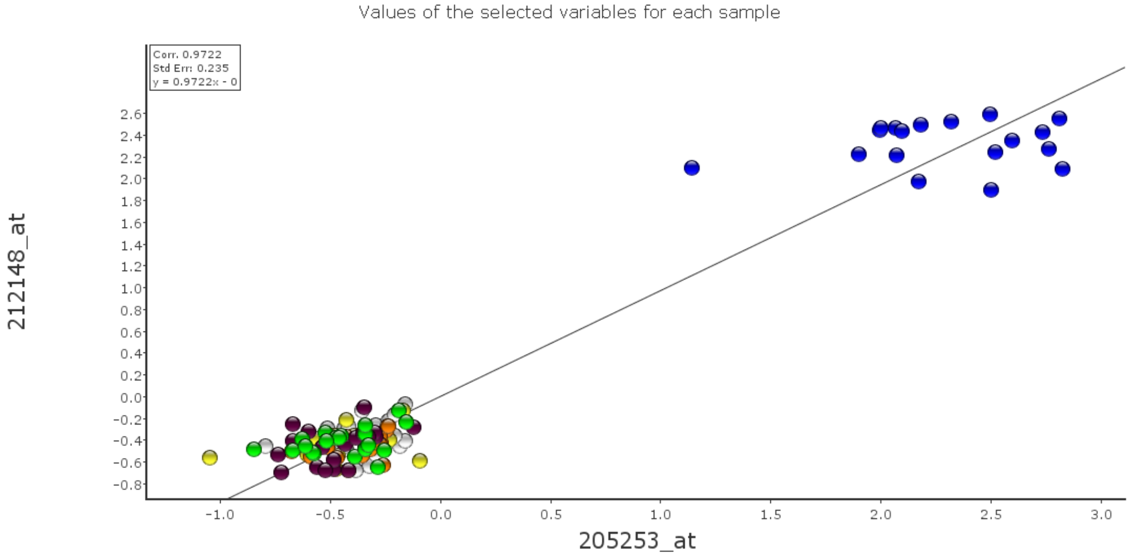


Scatter plots

The Scatter plots are very flexible and they can be populated with data in many different ways. The first example is a Sample Scatter plot.

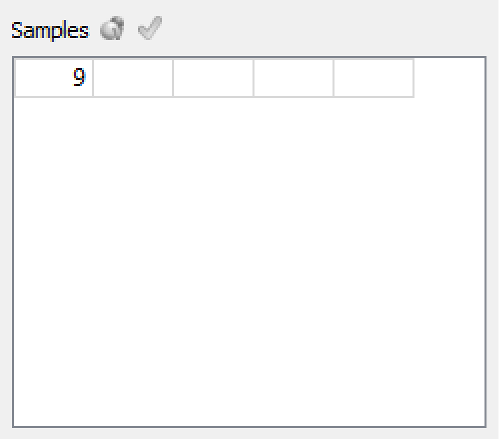
- Select plot type **Scatter** and mode **Sample** in the **Method Tab**.
- In the **Data Tab**, uncollapse the data. Change Variable identifier to probeset ID
- Select the **X Axis** tool in the **Tool Box** and select the *first* variable in the **Search list**
- Select the **Y Axis** tool in the **Tool Box** and select the *second* variable in the **Search list**
- Select the **View tab** and Color the samples by the annotation **"**Leukemia Subtype"

This should give you the following plot. One variable on each axis and all active samples plotted. Again the "E2A-PBX1" group (Blue) is separate in the plot since that is the behavior of the variables in the current **Search** list.

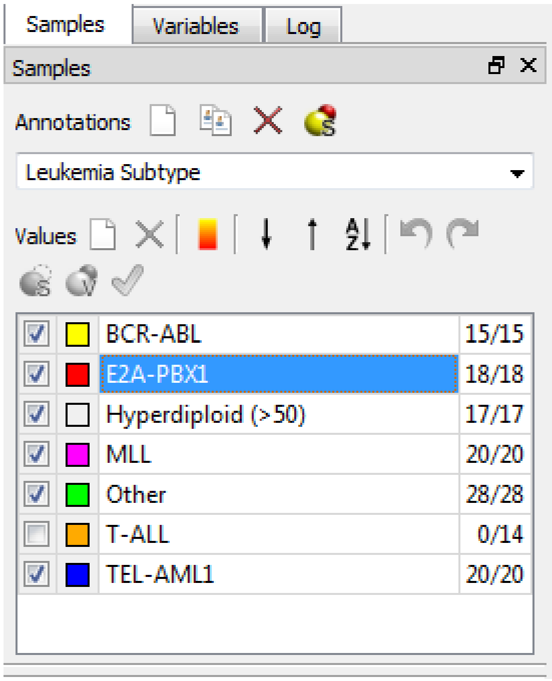


The second example is a **Variable Scatter** plot.

- Select plot type **Scatter** and mode **Variable** in the **Method Tab**.
- Select the **Sample Window** to see the sample annotations. Check that the annotation "Leukemia subtype" is visible.
- Select the **X Axis** tool in the **Tool Box** and select the *first* sample in the Sample Table (note that the numbers might be different).



- Select the **Y Axis** tool in the **Tool Box** and select the E2A-PBX1 group in the Sample Window.

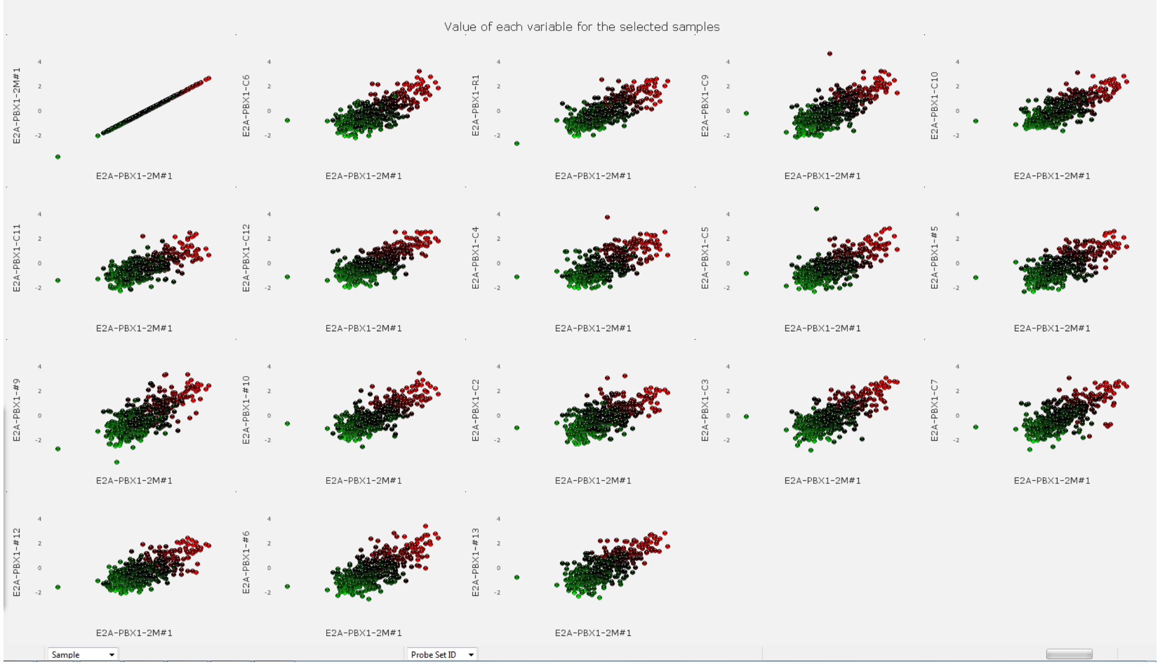


Leukemia Subtype

☑	🟨	BCR-ABL	15/15
☑	🟥	E2A-PBX1	18/18
☑	□	Hyperdiploid (>50)	17/17
☑	🟪	MLL	20/20
☑	🟩	Other	28/28
□	🟧	T-ALL	0/14
☑	🟦	TEL-AML1	20/20

- Select the **View tab** and select to Color the variables by the annotation "Leukemia Subtype".

This should give you the following plot. On the X-axis is the Sample named "E2A-PBX1-2M#1" in all subplots. On the respective Y-axis are the samples in the E2A-PBX1 group. Each point in the plot represents one of the 470 active variables.



Exporting images, animations and other data

You can at any time during an ongoing analysis export an image or an animation from QOE.
You do this by selecting **File > Export > Image** or **File > Export > Video,** and then supplying the name and other characteristics of the exported file.

You can also export principal components, sample distances, annotations and other useful data for downstream analysis, see **File > Export.**

In the **GSEA Workbench** there are two separate export functions: export List and Results. The List export transfers a copy of the selected lists to the QOE main program list interface.

In the **GO Browser** the Export List transfers the search results to the QOE main program list interface.

Also note that it is possible, at any point in an analysis, to save the complete current state of QOE by using the Log function**.** You can then return to that specific point in the analysis, by opening the corresponding log file in QOE, and select the specific log point.

Further Help

This tutorial has only covered a part of the functionality in QOE. For a more comprehensive coverage see the **Reference Manual** supplied in the **Help Menu**. Additionally you find a broad suite of on-line documentation at www.qlucore.com/documentation.aspx

Bibliography

[Ross2003] M.E. Ross et. al. *Classification of pediatric acute lymphoblastic leukemia by gene expression profiling.* Blood 15 October 2003, Vol 102, No 8, pp 2951-2959.

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 intended for research use only.

Trademark List

Windows Vista, Windows 7 and Windows 8 are trademarks of Microsoft.
Affymetrix is a trade mark of Affymetrix.
Agilent is a trade mark of Agilent.

References

GSEA: Subramanian, Tamayo, et al. 2005 Proc Natl Acad Sci U S A 102(43):15545-50.
GO: The Gene Ontology Consortium. "Gene ontology: tool for the unification of biology." Nat. Genet.. May 2000;25(1):25-9.
R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Appendix: An introduction to the statistical concepts

An introduction to statistical hypothesis testing

Statistical hypothesis testing is all about making decisions concerning one or more *populations*, using information provided by *sampling* from these populations. Before we begin, it is important that the populations that we are interested in are carefully defined and that the data sets obtained from the sampling are representative of these populations. For example, say that we are interested in testing if men are, on average, taller than women, and that we measure 100 randomly selected men and 100 randomly selected women from a specific geographic region. Can we then use the results from the statistical test to say something about the average heights in the worldwide populations of men and women? If there is a regional effect, the results may only be correct for the populations of men and women in the region we studied. The ability to make decisions about the populations using only the sampled information is important since it is often not feasible to study the entire populations (if we could, there would be no need for statistical tests). The drawback of this approach is, of course, that since we do not study the entire population, we can never draw conclusions about it with 100% certainty. The statistical hypothesis testing framework allows us to handle this uncertainty in a formalized manner.

What is a hypothesis?

A statistical hypothesis is a statement concerning the *populations* of interest. In the general hypothesis testing framework, we have a *null hypothesis* (H_0) and an *alternative hypothesis* (H_a). The null hypothesis often represents a state of "no effect". In the height example above, the hypotheses can be

- H_0: there is no difference between the average heights of men and women
- H_a: there is a difference between the average heights of men and women

To formulate the statistical hypotheses, we first need to formulate the *biological* hypotheses. In our example, the biological null hypothesis can be that the sex has no effect on the height of the individual, and the alternative hypothesis is that the sex indeed has an effect on the height. Then, we must formulate the biological hypotheses in terms of well-defined, measurable quantities (like the average height above). The choice of statistical test depends partly on which characteristic of the populations that we are interested in comparing, that is, how we make the translation from biological to statistical hypotheses.

What is a p-value?

The results of statistical hypothesis tests are often represented by means of p-values. To get from the observed values of our variable to a p-value, we first need to

construct a *test statistic*. The test statistic provides a numerical summary of the sample data and is designed to capture the effect that we are interested in studying. In principle, we can think of a lot of test statistics that would capture a given effect. The reason for selecting a particular statistic is often that under some assumptions regarding the underlying populations, we can calculate theoretically how the statistic would be distributed if the null hypothesis was indeed true. Then, we can compare the value that we computed from our sampled data to this distribution and say *how likely it would be to get a test statistic value that is as or more extreme than the observed one, given that the null hypothesis is true*. This probability is precisely the definition of the *p-value*.
In the example above, a small p-value would imply that it would be very unlikely to obtain a value of the test statistic that is as or more extreme as the one we have computed from our samples, if there was really no difference between the average heights of men and women in the population. Conversely, a large p-value means that it is quite likely to obtain such an extreme value *even* if the null hypothesis is true. Thus, in the latter case there would not be any significant evidence for a sex effect on height since apparently, we could very well have observed the heights obtained in our experiment even if there was no difference between the average heights in the female and male populations.

If the computed p-value is below a pre-specified significance threshold (by far, the most common significance threshold is 0.05) we *reject* the null hypothesis. Conversely, if the p-value is above the significance threshold, we *do not reject* the null hypothesis. A few things are worth noting (see also the article by Goodman[16] for a more extensive discussion):

- The p-value does *not* tell you how likely it is that the null hypothesis is true. Similarly, it does not tell you how likely it is that the alternative hypothesis is true.
- If you cannot reject the null hypothesis, you have *not* proven that the null hypothesis is true, but simply that the current data set did not provide enough evidence to reject it.
- There is nothing "magical" with the significance level of 0.05. In the early years of hypothesis testing, tables were used to determine rejection regions for the test statistic. These regions were pre-computed and tabulated for a few p-value thresholds only. However, today computers easily give the exact p-value and thus the actual values should be reported instead of just, e.g., p<0.05. Having the actual p-values is also necessary to compute corrected p-values (q-values, see below).

[16] Goodman, S: A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology 45:135-140 (2008).

What is a q-value?

Using p-values to interpret the result of a statistical test works fine if we only perform one test (that is, if we have only one variable in our data set). Once the number of tests increases, the usefulness of the p-value as a measure of significance decreases. To see why, assume that we have 10,000 variables in our data set, and that the null hypothesis is true for each and every one of them. Now apply a statistical test to each variable. Owing to the definition of the p-value, we expect 5% of the variables to give values of the test statistic that are more extreme than what is required to reject the null hypothesis at the 0.05 significance level. In this particular example, we would thus have approximately 0.05*10,000 = 500 variables with p-values below 0.05, even though the null hypothesis is actually true for all the variables! These are called false discoveries, or false positives. If there are indeed some variables in the data for which there is a true difference, they will be mixed with false positives.

An alternative interpretation, suitable for the situation where multiple tests are performed, is given by the *false discovery rate* (FDR). The FDR is the expected fraction of false discoveries among all the significant test results. It is possible to compute a corrected p-value, or a *q-value*, for each variable. The q-value is the FDR analog of the conventional p-value. For a given variable (say, with p-value p*), the q-value estimates the fraction of false discoveries among all variables with p-values below p*. Note that the q-value dQOEs *not* give the probability that the actual variable is a false positive. Hence, it dQOEs not tell you *which* variables are most likely to be false discoveries.
To get a feeling for false discovery rates, imagine taking all the computed p-values and lining them up, ordered in increasing order. Setting a significance cutoff now means to decide on a threshold (for example a p-value of 0.05), and considering all p-values below that threshold to represent "discoveries". The false discovery rate is the expected fraction of false discoveries among these, that is, the fraction of the discoveries for which the null hypothesis is really true (recall that the null hypothesis is defined in terms of the *population* parameters). A false discovery is thus a variable that obtains a low p-value just by random chance, without any true underlying signal in the population. One way to reduce the number of false discoveries would be to push the significance cutoff closer to zero. However, since the false discoveries are mixed with the true discoveries (the ones for which there is really an effect on the population level), this would also exclude many of the true discoveries. In other words, we often need to allow for a few potentially false discoveries to get to the true ones.

The q-values calculated by Qlucore Omics Explorer can be used in different ways. One approach is to decide which expected fraction of false discoveries that one is willing to accept and then set the desired q-value threshold, in the Statistics toolbox, to this fraction. Among the variables remaining after this procedure, you can expect the fraction of false discoveries to not exceed the specified fraction (note that it may happen that no variables remain after this procedure, if the desired false discovery rate limit is too stringent). What is an acceptable level of false discoveries of course depends heavily on the specific application, but 10% (i.e. a q-value threshold of 0.1) is reasonable in many cases. A second approach to using the q-values is to decide on the significance cutoff in another way (for example, based on p-values as illustrated above). The largest q-value among the remaining variables can then be used as an estimate of the fraction of false discoveries among these. Keep in mind that the q-values, just as the p-values, are linked to a specific test and thus check that the settings in the Statistics toolbox agree with the test that you wish to perform.

Should I use one-sided or two-sided tests?

The choice of one-sided or two-sided (sometimes one-tailed or two-tailed) tests comes down to the formulation of your hypotheses (and hence, the choice should be made before the test is applied). A one-sided test assumes that only deviations in one, pre-specified, direction are interesting (i.e., corresponding to the alternative hypothesis).

In most situations, a two-sided test is arguably the most appropriate, and the use of a one-sided test generally requires substantial motivation. Assume, for example, that we are trying a new drug, and we want to compare it to the conventional treatment. In this situation, we might assume that the new drug will perform better than the old treatment, and thus use a one-sided test. However, this means that even if the new drug turns out to perform much worse than the old one, all we can say from the one-sided test is that we cannot reject the null hypothesis that it is equally good or worse than the old treatment. Clearly, it may very well be of interest to know whether the new drug is in fact significantly worse than the old one, even if this result is unexpected.

What is a t-test and when should I use it?

A (two-sample) t-test is designed to compare the mean values of a variable between two populations. The t-test is commonly used in many practical situations and, although its validity depends on some underlying assumptions, it is often fairly robust to deviations from these. In particular this is true if the number of samples is large and equally distributed between the two sample groups.

The t-test assumes that the samples are collected independently. If, for example, all samples from one population were taken from the same subject, the variance may be seriously underestimated which leads to flawed inference. You can in general not test or correct for this once the data have been collected.

Another underlying assumption of the t-test is that the variable is normally distributed in each population. However, for large sample group sizes (more than, say, 20-30 observations per group), this assumption is not critical and also for smaller sample group sizes the t-test is often reasonably robust against deviations from normality.

You can check the normality assumption graphically in Qlucore Omics Explorer using box plots, which should be symmetric (i.e., with the median approximately in the middle of the box, with approximately equally long whiskers on each side). Unfortunately, it can be difficult to tell from a small set of samples whether the normality assumptions are satisfied or not (and if we have many samples, we saw that the assumption may not be that critical). Another possibility is to use previous experience, which may have shown that the variable under consideration is likely to be approximately normally distributed.
The robustness of the t-test also depends on the relative sample group sizes in our data set and whether or not the variances are equal between the two populations. The t-test employed by Qlucore Omics Explorer assumes equal variance between the populations. This assumption can be checked graphically using scatter plots of the variables. If the number of samples in both groups are equal, the t-test is fairly robust against violations of this assumption. See figures in the section on ANOVA below for illustrations of how to check the assumptions of normality and equality of variances graphically, using the tools provided within Qlucore Omics Explorer.

If the data shows considerable non-normality and/or inequality of variances, a transformation may be useful. For example, if the values are skewed to the right, a logarithmic transformation may provide a distribution that is closer to normal. One example of this is gene expression data obtained from microarrays, which is often assumed to follow a normal distribution after a logarithmic transformation. One should, however, note that transformations change the scale of the values and may in some cases make the results more difficult to interpret.

A *paired* t-test is used when the data come in pairs, for example, if each subject has been given both of two compared treatments. Applying a paired t-test can increase the power to detect differences between groups by accounting for individual differences between subjects.

What is ANOVA and when should I use it?

ANOVA (analysis of variances) is a generalization of the t-test, allowing comparison of more than two population means. We can also invoke several predictors and take covariates into account. The two-sided t-test is identical to a so called one-way ANOVA (an ANOVA with one predictor) where the predictor has two categories. When comparing the mean values of more than two groups, the null hypothesis is that all population means are equal, and the alternative hypothesis is that at least one mean differs from the rest. Hence, the ANOVA dQOEs not immediately tell us *which* of the means are different. Moreover, directed (one-sided) tests do not make sense when more than two means are compared. For a more comprehensive overview of different ANOVA models, see the document "How to use ANOVA" which is available from www.qlucore.com.

As a generalization of the t-test, ANOVA is built on the same assumptions of independence within groups, normality and equality of variances. As for the t-test, the ANOVA is fairly robust against deviations, especially for equal group sizes (and preferably large number of samples). Figures 1 and 2 shows how the assumptions of normality and equality of variances can be graphically checked within Qlucore Omics Explorer.

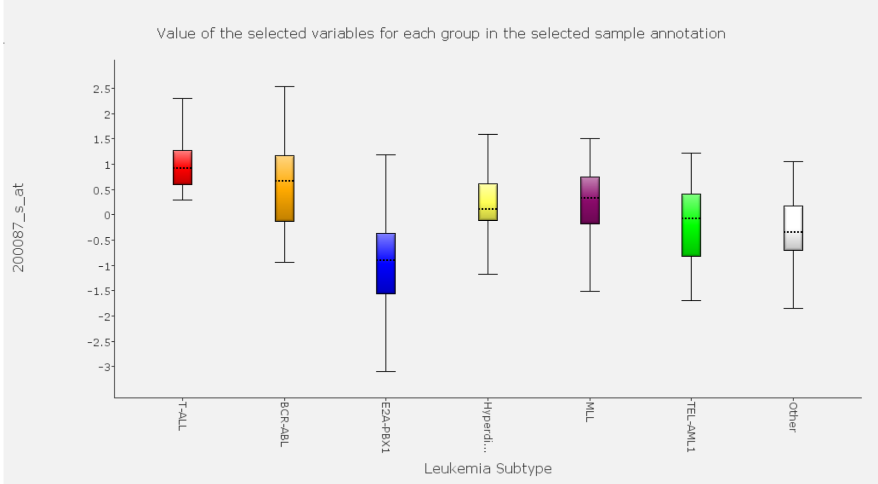


Figure 1. A variable which is approximately normally distributed within each sample group.

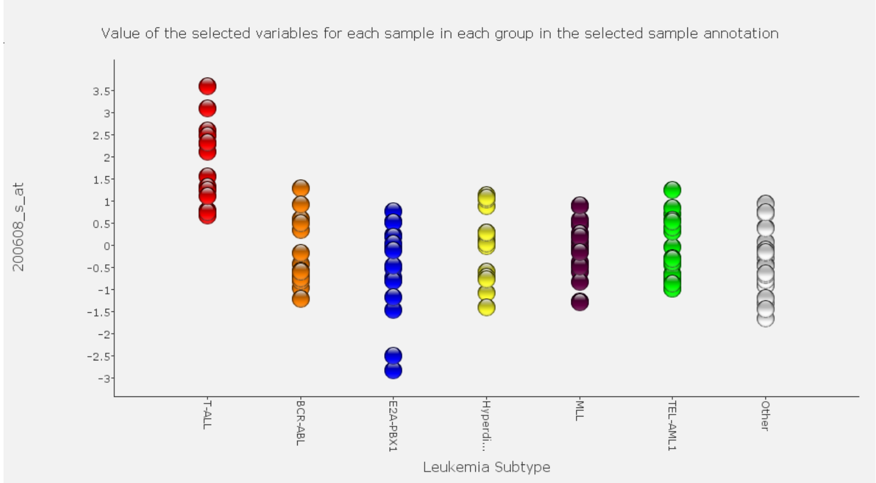


Figure 2. A variable which shows similar variances (but different means) in the different sample groups.

When should I use variance filtering?

Variance filtering can be used as a way to reduce the number of variables in a data set in an unsupervised way (i.e., without using any sample annotations). With current experimental techniques it is easy to collect data for a huge number of variables simultaneously, and the variables are generally not explicitly chosen based on prior

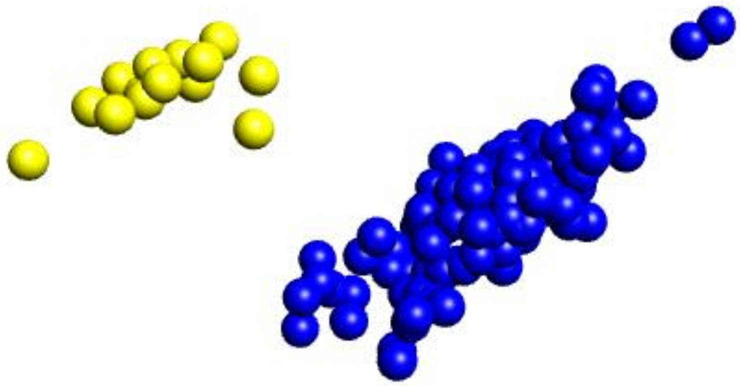
assumptions of "interestingness". Hence, it is conceivable that a large part of the variables add nothing but noise to the analysis and we may want to remove these variables to explore the rest of the data in more detail. One way to identify potentially uninteresting variables is by means of their variance. A variable which is almost constant across all the observations has a low variance, and we can remove such variables by means of the variance slider in the Statistics toolbox.
It has also been suggested that variance filtering may be useful to increase the power of detecting differentially expressed genes with t-tests or ANOVA. The rationale behind this is that the variance filtering reduces the number of variables and therefore makes the correction for multiple testing (see the discussion on q-values above) less impeding.

The appropriate amount of variance filtering depends heavily on your data set and the goal of the analysis. If the data set has not been pre-filtered before it is loaded into Qlucore Omics Explorer, filtering out more than half of the variables is conceivable. If the data set has been pre-filtered, considerable less variance filtering may be needed.

Appendix: PCA plots

The basic meaning of the PCA plot of any multidimensional data in QOE is that data points that are similar are also presented close together in the generated plots.
The PCA operation is characterized by the feature that it preserves as much of the originally available information as possible in the generated three-dimensional plots. The information content is then measured by the statistical variance in the data when applying PCA.

In the picture below, the yellow group consists of samples that are similar to each other and that are different from the blue samples.

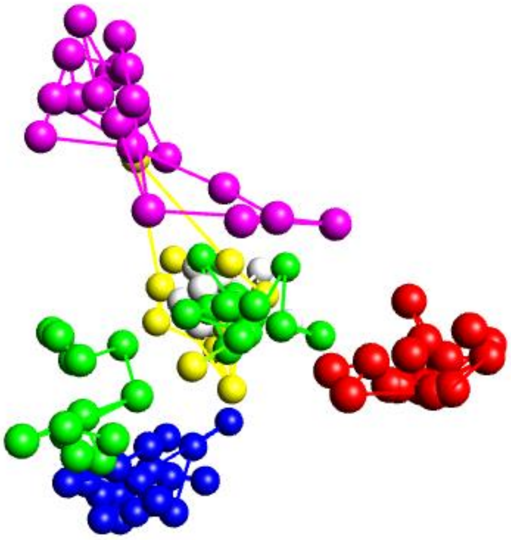


What is the statistical significance of a PCA plot. Can I trust it?
The PCA operation in QOE does not make any assumptions regarding your data. If you can see structure and patterns visible on the computer screen it is then because that structure is present. Some statistical methods provided in QOE (such as ANOVA) may create patterns even from random data. These patterns are then, with very high probability, not statistically stable and you must look at the **statistical significance** of the structures you discover. QOE comes with several available tools for controlling **statistical significance**. They include **cross validation** (leaving one or several data samples out), **randomization** or **permutation** tests. QOE also provides **p-values and q-values** for the chosen statistical methods, making it easy to dynamically check the statistical significance of the structures you discover.

Can PCA miss structure and patterns?
The PCA operation is used to reduce dimension and hence there is in general a loss of information in the three dimensional presentations. The PCA operation nevertheless is

a stable and in a certain sense optimal method for dimensionality reduction and by using the flexibility of the Dynamic PCA functionality in QOE you minimize the risk of missing important structures.

The use of graphs and nonlinear methods such as ISOMAP provided in QOE is also a way to minimize the risk of missing vital information concerning your data. In the picture below you can for instance, with the use of graphs, see that the green group in fact consists of two different subgroups. This fact would have been hard to discern without the support of the graphs present in the plot.



Appendix: Clustering and classification

Clustering and classification algorithms are examples of *machine learning* methods, where the computer uses a specified procedure to “learn” something about the data. The main difference between them is that clustering is *unsupervised*, while classification is an example of a *supervised* approach. These two types of methods are used in different contexts and for different purposes. Unsupervised methods do not use any external information (annotations, such as disease status or other traits) about the objects to be analyzed, but rather try to find dominating structure or patterns in the data, patterns that can then be interpreted by the researcher. Clustering is an example of an unsupervised method since the goal is to partition the objects into subgroups, without using any external annotation information. Also PCA is an example of an unsupervised method. Supervised methods like classification, on the other hand, typically aim at building models that predict or "explain" some pre-specified annotation, e.g., disease status or the response to a treatment. This annotation may or may not correspond to the main pattern(s) in the data. Given some data and a sample annotation, the aim is to build a model from the data that is able to predict the value of the sample annotation in a new sample for which we are only given the data.

Should I use an unsupervised or a supervised method?

It is important to recognize if the goal of a study requires a supervised or unsupervised approach. For example, if the goal is to build a model that can predict the disease status of a patient, we should use a supervised approach. Using an unsupervised approach like clustering or PCA will likely mix the signal that we are interested in with other, unrelated, signals and generate a worse predictor, unless the disease status is actually the dominating signal in the data. On the other hand, if the goal is to get an overview of a data set, to see which are the strongest patterns and whether the samples naturally partition into subgroups, an unsupervised method like clustering or PCA should be used.

An aspect to keep in mind when using supervised methods, especially on high-dimensional data sets, is that since we are explicitly searching for patterns that are associated with the annotation we want to predict, in the vast data space we will most certainly find something that can predict the annotation well in the current data set. However, this is not what we are ultimately interested in (since we already know the annotation values in this data set). Instead, we are interested in knowing whether the derived model can generalize, i.e., predict the value of the annotation in an independent data set that the model has not seen before, and where we may have only the data, but no information about the annotation. Thus, supervised models must always be validated in independent data sets, and a good predictive performance in the training data does not provide any evidence that the model is good. This is usually less important for unsupervised methods, which are more often used to summarize, explore and describe a given data set.

Clustering

Qlucore Omics Explorer offers two types of clustering methods: hierarchical clustering (combined with heatmaps) and k-means clustering. Both are used for the same purpose: to find subgroups among the samples, such that samples within one group are more “similar” to each other than samples belonging to different groups, where “similar” can be formally defined in various ways. The difference is that the hierarchical clustering builds a “cluster tree” (or dendrogram), which organizes the samples hierarchically but does not directly divide them into clusters, while the k-means clustering partitions the samples into a pre-defined number of groups.

Practical situations where one would like to use a clustering approach are e.g.:

- to evaluate whether there are subtypes of a particular disease, i.e., if the samples group into different clusters based on some measured data. These clusters may represent different disease types, which have different prognosis and behavior.

- to explore a data set and look for artifacts. This can be done by clustering the data and examining whether the obtained clusters are associated with the signal(s) or interest, or rather with spurious ones such as batch effects or other technical artifacts.

Classification
Classification models consist of two parts: the variables that are used and a rule to combine the values of these variables in order to obtain a predicted value of a given sample annotation. Both are important, and are usually determined together. Qlucore Omics Explorer provides several ways of building a classification model, such as random trees, support vector machines and k-nearest neighbor algorithms.

Practical situations where one would like to use a classification approach are e.g.:

- to build a model that can use gene expression values to predict the prognosis of a cancer patient
- to build a model that can assign a sample to one of several previously defined disease subtypes, based on some observed biomedical data

As noted above, it is important that a predictive model is evaluated on independent data, and not on the same data where it was built. *Overfitting* refers to the situation where a model is “too specifically adapted” to a given data set and does not generalize to other data sets. Usually this is a sign that the model has adapted too much to the random noise in the training data set, in its strive to build a model that fits well in this data. The noise in an independent data set will likely be different, and then the model may not work anymore.

Cross-validation is a technique that can be used to estimate a model’s expected performance based on a single data set. The underlying idea is to subdivide the entire data set into a training and test set (multiple times), build the model on the training part and evaluate the performance on the test part (which was not used to build the model).
The word classification is usually used to describe predictive modeling where the sample annotation is categorical. To predict a numeric/continuous annotation, one typically uses regression.

The three different classifier models have different characteristics:

- kNN: classifies a sample to the majority class among its k nearest neighbors.
- Support Vector Machines (SVM): attempts to construct a hyperplane (a high-dimensional linear boundary) that separates the samples in the two compared classes as well as possible. Each new sample is then classified based on which side of the boundary it falls.

- Random Trees: builds a large set of classification trees. Each tree consists of a (typically small) number of successive branching points, and in each of these a sample is assigned to the “left” or “right” branch depending on the value of one variable. The leaves of each tree correspond to different classes, and the class assignment of a sample is determined by the leaf where it ends up. To improve performance and stability, multiple trees are built (based on different subsets of the variables) and the final class assignment of a sample is determined by summarizing the classifications from all the trees.