How to work with multiple data sets

INTRODUCTION

There are many situations when it might be important to work with multiple data sets that are related in various ways.

One example is when we measure multiple types of entities for the same samples and we then are interested in finding out for example whether a grouping of the samples that is found in one of the data sets is present also in the other data set.

Another example is when results generated from one data set need to be validated using another, independent data set. For example, a useful biomarker candidate for a specific disease, which can be suggested from exploratory analysis of some data set, should be able to single out the patients with the disease in any similar data set.

Qlucore Omics Explorer (OE) offers straightforward ways of transferring information about samples or variables between multiple data sets. The analysis is facilitated since in OE, you can have multiple data sets open in parallel. Variable information is generally transferred using variable lists, while sample information is transferred by means of sample annotations.

TERMINOLOGY

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

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

We use annotations to denote descriptions of samples or variables. One sample or variable can be described by one or many annotations.

Normally a **data set** is described by a matrix where the columns represent samples and the rows represent variables. We assume that we have two data sets. One of the data sets, called the training data set, is used to generate hypotheses and find potentially interesting structures. The other data set, called the validation data set, is used to validate, confirm or further explore the generated hypotheses from the training data set.

COMPARING (VALIDATING) RESULTS IN THE SAMPLE DIMENSION

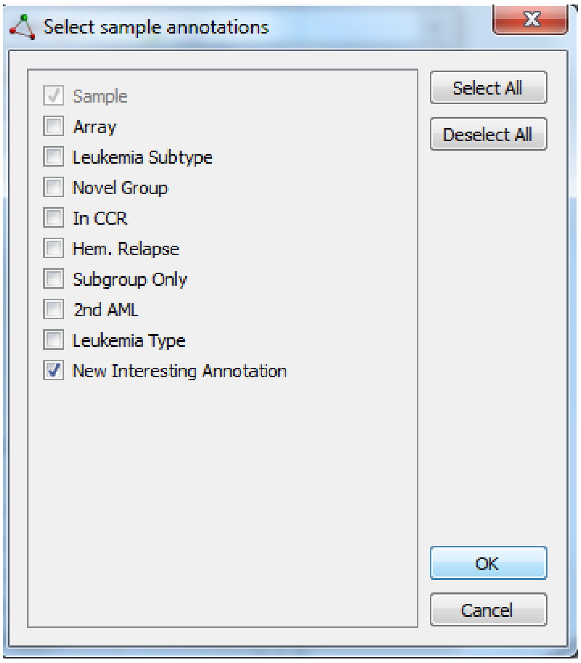
To see whether a sample grouping found in the training data set can be seen also in the validation data set, the following steps could be used:

1. Open both your training and validation data set(s) in Qlucore. You can have multiple data sets open at the same time. In the screen below you have the training data set to the right and the validation data set to the left.

2. Explore the training data set in order to find a potentially informative grouping(s) of the samples. Create new sample annotation(s) that captures this grouping. See for example the document on How to find structure and patterns for more details about this procedure. In the screen the new interesting annotation is used to color the training data set in red and yellow were we are interested in the red group.

3. Export the new sample annotation using the File – Export – Sample annotations option.

4. Import the saved sample annotations into the validation data set. In the picture below the Select sample annotation dialog is open after the use of File – Import – Sample annotation. In the example we select the “New Interesting Annotation” for import into the validation data set.



Select sample annotations

☑ Sample
☐ Array
☐ Leukemia Subtype
☐ Novel Group
☐ In CCR
☐ Hem. Relapse
☐ Subgroup Only
☐ 2nd AML
☐ Leukemia Type
☑ New Interesting Annotation

Figure 1

5. Color the samples in the validation data set according to the imported annotation to see whether the same grouping is visible also in the validation data set. In the plot below, this has been done. We can clearly identify that the samples split according to the value of the New Interesting Annotation also in the validation data set.

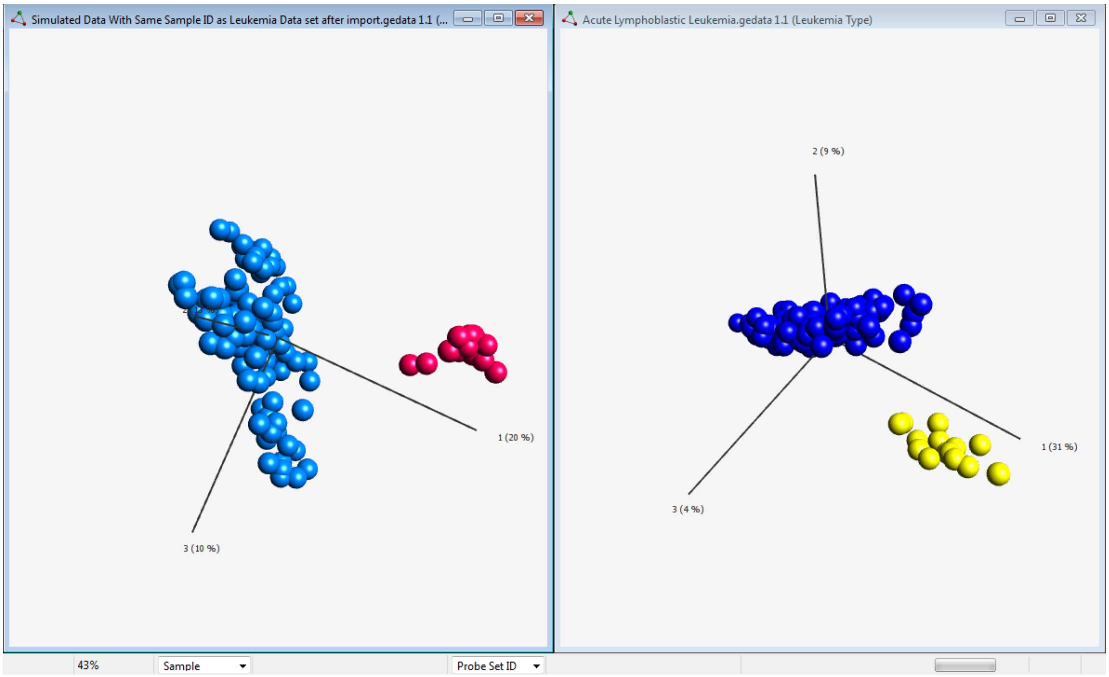

Simulated Data With Same Sample ID as Leukemia Data set after import.gedata 1.1 (...
Acute Lymphoblastic Leukemia.gedata 1.1 (Leukemia Type)
1 (20 %)
3 (10 %)
2 (9 %)
1 (31 %)
3 (4 %)
43%
Sample
Probe Set ID

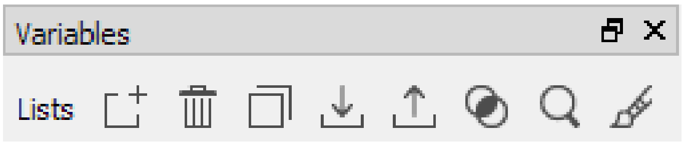
Figure 2

COMPARING(VALIDATING) RESULTS IN THE VARIABLES DIMENSION

To find potential variables of interest in the training data set and validating them in the validation data set, the following steps could be used:

1. Load both the training data set and the validation data set into OE.

2. Make sure that the variable annotations that are used as the unique variable identifier (ID) in the two data sets match. This is necessary in order to transfer variable information between the data sets. The annotation used as variable ID can be changed using the ID button in the variable tab. [1]

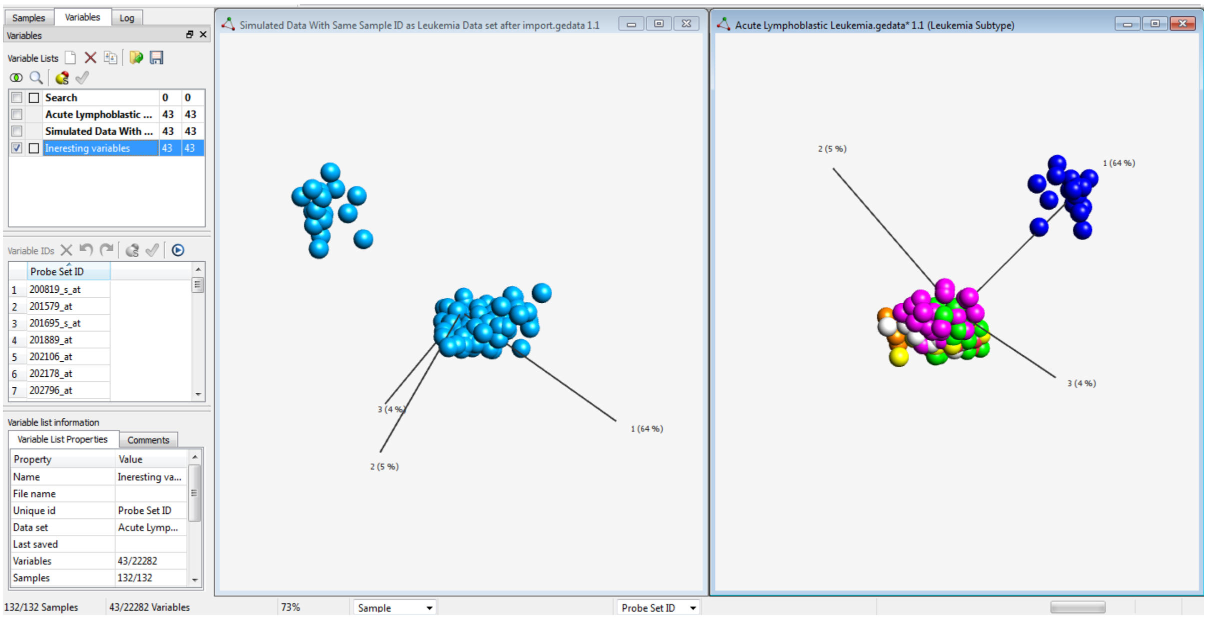


3. Generate a potentially interesting variable signature in the training data set. This can be done for example by using a statistical test to find variables that are significantly associated to a given sample annotation, or by using the Search tool to search the available variable annotations and extract variables of interest (see the How to use search document for more details about this). Once the variables of interest have been found, they are contained in the active variable list in the training set (the list has the same name as the training data set). See

[1] If your data set do not have matching variable IDs you will have to translate the list of variables generated in the training data set to a list of identifiers that match those of your validation data set.

the screen below were the plot to the right is the training data set and we have created a list of interesting genes that separates the blue group. The list can be found in the variable panel to the upper left.

4. Switch to the validation data set. To study the behavior of the generated variable signature in the validation data set, select the active variable list from the training set as the input to the validation data set by checking the check-box before the variable list. The screen below shows the results. Apparently, the generated variable signature distinguishes two groups of samples in the validation data set. The result is shown in Figure 3.


Samples
Variables
Log
Variables
Variable Lists
Variable IDs
Simulated Data With Same Sample ID as Leukemia Data set after import.gedata 1.1
3 (4 %)
2 (5 %)
1 (64 %)
Acute Lymphoblastic Leukemia.gedata* 1.1 (Leukemia Subtype)
2 (5 %)
1 (64 %)
3 (4 %)
73%
Sample
Probe Set ID

□	□	Search	0	0
□		Acute Lymphoblastic ...	43	43
□		Simulated Data With ...	43	43
☑	□	Inerestina variables	43	43

	Probe Set ID
1	200819_s_at
2	201579_at
3	201695_s_at
4	201889_at
5	202106_at
6	202178_at
7	202796_at

Variable list information

Variable List Properties | Comments

Property	Value
Name	Ineresting va...
File name	
Unique id	Probe Set ID
Data set	Acute Lymp...
Last saved	
Variables	43/22282
Samples	132/132

132/132 Samples | 43/22282 Variables

Figure 3

As an alternative to step 4, copy the active variable list from the training data set and save the copy of the list. You can then import the saved variable list and use it as the input in the validation data set, even after the training set has been closed.

DISCLAIMER

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

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

Qlucore Omics Explorer is only intended for research purposes.