How to import data

Contents

1. Introduction 2
2. Supported data formats/arrays 2
3. Aligned BAM files 3
4. How to load and open files 3
5. Affymetrix files 4
   5.1 Affymetrix CEL files (.cel) 4
   5.2 Affymetrix Probe Set Files (.chp) 4
6. Agilent Files 5
   6.1 Agilent Text Files 5
   6.2 Agilent GeneView files (.txt) 5
7. Import Wizard 5
8. Qlucore Data Files (.gedata) 6
   8.1 How to create a Qlucore Data File 6
   8.1.1 Formatting 7
   8.1.2 Layout 7
   8.1.3 Transposed data 9
9. Simple Data File 9
   9.1 Compact text file (.csv) 9
10. Load annotations 10
   10.1 From a .txt or .csv file 10
   10.2 Create a sample/variable annotation text file 11
   10.3 From NetAffx 12
11. R users 12
12. Templates 12
13. NGS Data 13

Terminology 13

Decimal point and decimal comma 13

Disclaimer 13

Trademark List 13

1. INTRODUCTION

With the introduction of Qlucore Omics Explorer version 3.3 the program can be configured with separate modules. This document covers the import functionality of the Qlucore Omics Explorer base module.

2. SUPPORTED DATA FORMATS/ARRAYS

This section applies to experiment data. Other sections will cover import of annotations (clinical information) and system biology related information such as gene sets and pathways.

The Qlucore Omics Explorer (QOE) base module supports direct import of experiment data including normalization for:

- Aligned BAM files with RNA-seq data
- Affymetrix 3" and WT arrays.
- Agilent mRNA arrays and Agilent microRNA arrays

For Illumina array data the recommended workflow is to normalize data with the GenomeStudio or BeadStudio software and then use the Wizard in Qlucore Omics Explorer.

For data generated with other arrays/instruments or resulting from other type of sources, Qlucore offers a wide range of import alternatives. If data is stored in a .txt, .csv or .tsv file the Import Wizard, see chapter 5, is a useful tool for efficient data import**. If you are uncertain of how your data is structured the Wizard is normally the best way to import data.**

In total 10 different data set file formats are supported in the QOE base module:

- BAM files (.bam)
- Affymetrix CEL files (.cel)
- Affymetrix Probe Set Files (.chp)
- Agilent Text Files (.txt)
- Agilent Gene View files (.txt)
- Simple Data Files (.txt)
- Qlucore Data Files (.gedata)
- GEO Data Sets (.soft and .soft.gz) [1]
- GEO Series Matrix (.txt and .txt.gz) [2]
- Compact Text Files (.txt; .csv)
- BioArray Software Environment Files (.base)

[1], [2] Note that it is only Data Sets from Gene Expression Omnibus in these formats that are supported. Not just any *.soft or *.soft.gz file.

3. ALIGNED BAM FILES

You can import and normalize RNA-seq data in the form of an aligned BAM file.

Select the **File → Open BAM** files menu item.

Use the Add button to select individual files.

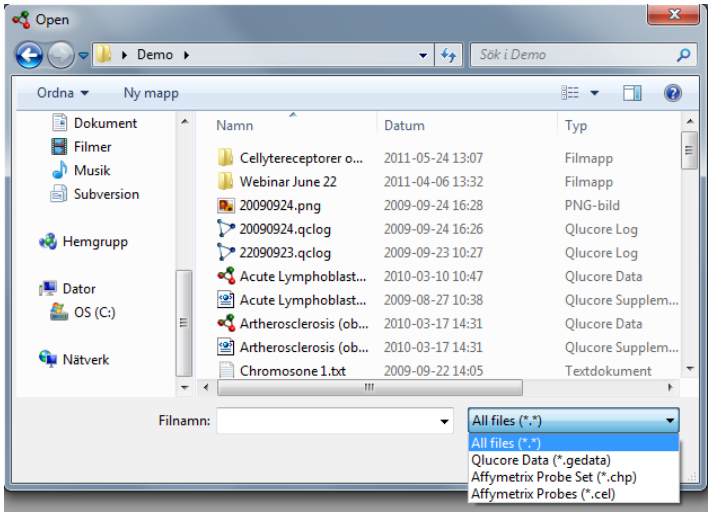
Or use the Add Folder button to select all BAM files in a folder recursively.

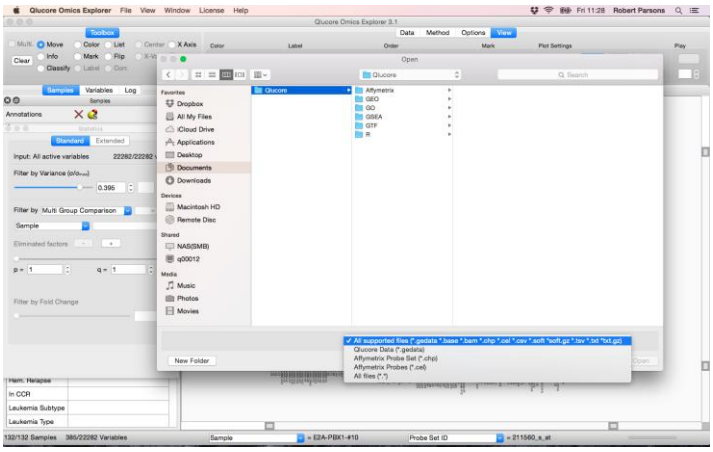
Press OK to start loading the selected files. This can take a while.

A GTF (Gene Transfer Format) file is needed to describe where the genes are located in the reference genome. file. GTF files can be downloaded from ftp://ftp.ensembl.org/pub/current_gtf/.

4. HOW TO LOAD AND OPEN FILES

Files are loaded by using the File->Open menu. Then there are different filters that you can select. See the picture below. *Note that the picture might vary slightly depending on your operating system.*





It is also possible to load both variable and sample annotation files, see chapter 10.

5. AFFYMETRIX FILES

This applies to mRNA array based data. For miRNA data, only normalized data is accepted, and the suggested workflow is to import the normalized data using the Wizard, see section 6.

5.1 AFFYMETRIX CEL FILES (.CEL)

The import and normalization will normally include four steps

1. Select files
2. Select normalization method
3. Select which annotations to import
4. Inspect the QC-report

To start the import: Select File->Open and select the files you would like to open. To select multiple files, use the shift key on the keyboard.

The data import process includes downloading information about the array from Affymetrix server. You need to be connected to Internet if you do not have this data stored locally.

After a short while you are asked to select normalization method. Three methods are provided

- RMA
- RMA Sketch
- Plier

The next step is to download annotations from the Affymetrix NetAffx server [3] . You can select which annotations to include in your data set.

Finally, the QC-report is displayed.

In OE, normalization and summarization will be performed on gene (transcript level) both for 3" arrays and WT arrays.

5.2 AFFYMETRIX PROBE SET FILES (.CHP)

Select **File->Open** and select the files you would like to open. To select multiple files, use the shift key on the keyboard.

You will be asked if you would like to download annotation data from Affymetrix NetAffx server automatically[4]. When loading probe set annotations from Affymetrix Whole Transcript arrays, such as Human Gene 1.0 ST, Omics Explorer automatically creates simplified

---

[3],[4] This requires that the user is registered at the Thermo Fisher(Affymetrix) website. See chapter **Fel! Hittar inte referenskälla.** for more information about downloading this website.

annotations (Gene Symbol, Gene Title,…) in addition to the annotations in the Affymetrix annotation file. This makes it easier to interpret the results and to search for interesting genes.

6. AGILENT FILES

6.1 AGILENT TEXT FILES

Omics Explorer can import text files created with Agilent Feature Extraction Software. These files contain one array (sample) per file.

How to open Agilent text files

Select the **File→Open** menu item.

Select the Agilent Text File (*.txt) file filter.

Select the files and press Open. (Normally you select several files.)

In the Normalization Dialog, select a preprocessing scheme by pressing one of the four quick setting buttons.

- One-color mRNA
- One-color miRNA
- Two-color mRNA
- Two-color miRNA

Instead of using the above quick settings, you can select your own customized preprocessing. See the OE reference manual, section “Agilent .txt files”, for details.

6.2 AGILENT GENEVIEW FILES (.TXT)

Qlucore Omics Explorer can open Agilent GeneView files. These files contain miRNA expression levels. They are created with the Agilent Feature Extraction Software. Each file contains data from a single array. To select multiple files use the shift key on the keyboard.

Normalization

The data is normalized as recommended by Agilent:

- thresholding
- 2-logarithm
- percentile shift (optional, but recommended)

You can select the threshold value and the percentile. Recommended values are threshold = 1.0 and percentile = 75.

7. IMPORT WIZARD

The OE Import Wizard supports import of normalized data in many different formats. The starting point is that data is stored in a file with extension .txt, .csv or .tsv.

The samples can be in one or multiple files. If the samples are in multiple files the file format for all samples must be identical. You reach the Wizard by the menu **File->Open with Wizard.**

The Wizard guides you through 8 steps where you define how the layout of the data shall be interpreted and imported. The Wizard can for example handle

- Samples in multiple files
- Samples as columns and variables as rows, or the transposed version
- Various data separators [“,”, “;”, “tab” or other]
- Data on every ith column or every ith row.
- File header or not
- Data that shall not be imported
- Various indications of missing values [empty cell, n/a, NA,…]
- Annotations

If you are uncertain of how your data is structured, then the Wizard is the best option for data import.

Note: The Import Wizard is an automatic import method for very many different types of files. Make sure to verify that the import results in the expected number of samples and variables.

8. QLUCORE DATA FILES (.GEDATA)

The Qlucore Data file format has been designed to make it easy to create a file by copying and pasting in a spread sheet application. Data resulting from other tools and applications can easily be formatted and imported into Qlucore Omics Explorer using this file format. A Qlucore Data File contains a complete data frame, including sample and variable annotations. [5]

There are three important things to note when a file shall be created

1. It is a tab-separated text file.
2. The file shall have the file name extension .gedata.
3. Data shall be formatted as described below

8.1 HOW TO CREATE A QLUCORE DATA FILE

There are several ways to start the creation of the data file (.gedata).

a) If you have installed OE on your computer you can create a new data file template by right clicking on your desktop and selecting New and Qlucore Omics Explorer. This will generate a new file on your desktop. If your right click on the new file you can open the file with Excel or another spreadsheet application. Then you need to copy and paste data from your original data file into the new file. A guide on how to do this follows below in the section Formatting.

[5] A Qlucore data file does not contain information about the colors used to represent the values of a sample attribute in a plot, or the default variable annotation used when matching variable name lists with variable annotations in the data frame. That information is stored in supplementary data files.

b) Open your original data file and include rows and columns as described in the section Formatting.

8.1.1 FORMATTING

Copy and paste the data into the spreadsheet, following the layout below. You can use any spreadsheet application, such as Excel or OpenOffice. When you are done and have formatted data as below, go through the following three steps.

1. Select 'Save as' in the spread sheet application
2. Save the spread sheet as a tab delimited text file.
3. Rename the file by changing the file name extension from the default .txt to .gedata.

8.1.2 LAYOUT

The layout of the spreadsheet is best explained through an example. The data below contains

1. The measured data (cells are blue)
2. Sample annotation IDs: Array ID, Age, Sex (cells are pink)
3. Sample annotations (cells are yellow)
4. Variable annotation IDs: VarID and Symbol (cells are light blue).
5. Variable annotations (cells are green)

We use the following data frame:

Array ID	5301	5302	5303	5304
Age	34	22	47	41
Sex	Female	Male	Male	Female

VarID	**Symbol**
3140	MR1
622	BDH1
7551	ZNF3
1537	CYC1
961	CD47

2.28	-1.23		0.45
1.04	0	-0.03	0
-0.67	3.14	2.18	0.53
-1.34	2.34	-0.30	2.73
2.73	1.07	0.83	-1.52

In a spread sheet application, this data frame should be arranged as follows:

Qlucore	gedata	version 1.0						
4	samples	with	3	attributes				
5	variables	with	2	annotations				
		Array ID	5301	5302	5303	5304		
		Age	34	22	47	41		
		Sex	Female	Male	Male	Male		
varID	Symbol							
3140	MR1		2.28	-1.23		0.45		
622	BDH1		1.04	0	-0.03	0		
7551	ZNF3		-0.67	3.14	2.18	0.53		
1537	CYC1		-1.34	2.34	-0.3	2.73		
961	CD47		2.73	1.07	0.83	-1.52		

Always included

Note that the start of the file shall include the text and information as indicated by row 1 to 5 in the example above.

For your own data you need to adjust the number of samples, variables and annotations.

8.1.3 TRANSPOSED DATA

The example given above has samples as columns and variables as rows. For some data it is more natural to have the opposite, i.e. samples as rows and variables as columns.

Qlucore OE is capable to import .gedata files with data organized in this way if the word "transposed" is added to cell D1, see the table below. Note that the table is not complete.

Qlucore	gedata	version 1.0	transposed			
5	samples	with	2	attributes		
4	variables	with	3	annotations		
		Variable ID	101	102	103	104

9. SIMPLE DATA FILE

The simple data file is a tab separated text file with the extension .txt. The table below shows how data and identifiers should be organized. The rows are variables and the columns are samples. The first column should include a unique variable identifier (green cells) and the first row a unique sample identifier (yellow cells). Data is stored in the blue cells.

	5301	5302	5303	5304	
3140	2.28	-1.23		0.45	
622	1.04	0	-0.03	0	
7551	-0.67	3.14	2.18	0.53	
1537	-1.34	2.34	-0.3	2.73	
961	2.73	1.07	0.83	-1.52	

If there is a text in the upper left cell, that information will be used as the variable annotation header, otherwise the name will be *Variable ID*. The sample annotation header, corresponding to the sample annotation in the first row, will be the *Sample ID.*

Normally you would like to complement your data import with annotations, see chapter 10.

9.1 COMPACT TEXT FILE (.CSV)

Omics Explorer supports compact text files. These contain data, sample annotations and variable annotations for multiple samples in a single file. If compared to a ".gedata" file, see chapter 7, the compact text file is lacking a file header and has samples as rows. The layout of the files is described below. The last variable annotation (in the example below the second

row) will be used as variable id. Note that the name of this annotation is not included in the file; after loading the file the annotation will be called Variable Id.

The file can have either the extension ".txt" or ".csv". Both options work.

		Symbol	ENC1	CDK8	PEX7	SNN	DLX6
Array	Age	Gender	201314_at	204831_at	205420_at	218032_at	221289_at
5301	34	Female	2.28	1.04	-0.67	-1.34	2.73
5302	22	Male	-1.23	0	3.14	2.34	1.07
5303	47	Male	1.45	-0.03	2.18	-0.3	0.83
5304	41	Female	0.45	0	0.53	2.73	-1.52

How to open a compact text file

Select the **File** → **Open** menu item.

Select the file and press Open.

If the file contains only one variable annotation you will be asked to enter the index of the first data column (in the example above it is column 4).

10. LOAD ANNOTATIONS

Annotations are information about your samples and variables. It can for instance be clinical information about patients. The description below is a summary of options. More details are provided in the How to load annotations document and the section with the same name in the Documentation and Help manager in the program itself.

Note. To import annotations a data set, need to be loaded.

10.1 FROM A .TXT OR .CSV FILE

There are two options. Using the Annotation Wizard or importing a file of the format specified below.

Using the Annotation Wizard

Select the **File** → **Import** → **Sample Annotations via Wizard** or the **File** → **Import** → **Variable Annotations via Wizard** menu item.

The Annotation Wizard is launched and by following the instructions it is possible to import data from *.txt, '.csv or *.tsv files. The Wizard will assist you to pick out the data you need.

Direct From a .txt or .csv file

You can add sample and variable annotations to an existing data set by importing an annotation text file. These files are tab- or comma-delimited text files. The file name extension is .txt if the file is tab-delimited and .csv if the file is comma-delimited.

How to import a sample/variable annotation text file

Select the **File** → **Import** → **Sample Annotations** or the **File** → **Import** → **Variable Annotations** menu item.

Select the Tab-separated Text (*.txt) or the Comma-separated Text (*.csv) file filter.

Select the file and press Open.

Select the annotations you want to import and press OK.

When importing an annotation text file, the samples/variables in the file will be matched with the samples/variables in the data frame using the annotation in the first column of the annotation file and the sample/variable ID annotation in the data set (This can be changed in the Data tab). In this way the ordering of the rows in the annotation file and the data frame does not matter.

10.2 CREATE A SAMPLE/VARIABLE ANNOTATION TEXT FILE

Place the data in a spreadsheet, following the layout below. You can use any spreadsheet application, such as Excel or OpenOffice. Save the spreadsheet as a tab- or comma-delimited text file. Use the file name extension .txt if the file is tab-delimited and .csv if the file is comma-delimited. We recommend using tab-delimited text files.

The following layout should be used:

Probe Set	Transcript	Gene Symbol	Chromosome	Entrez Gene
1007_s_at	U48705	DDR1	chr6p21.3	780
1053_at	M87338	RFC2	chr7q11.23	5982
117_at	X51757	HSPA6	chr1q23	3310
121_at	X69699	PAX8	chr2q12-q14	7849
1255_g_at	L36861	GUCA1A	chr6p21.1	2978
1294_g_at	L13852	UBA7	chr3p21	7318

The first row contains the name of each annotation. Each of the remaining rows contains the value of each annotation for one sample/variable. The first column should contain an annotation that matches the sample/variable ID in the data set.

This table should be saved as a tab-separated text file with file name extension .txt or as a comma-separated text file with file name extension .csv.

10.3 FROM NETAFFX

You can download variable annotations for Affymetrix arrays from the Affymetrix NetAffx server and add them to the data set.

The downloaded annotation files will be stored on your computer, and can be used without having to download them from the server again. When you select to download annotations, you will be notified if updated annotations are available on the Affymetrix server.

When downloading probe set annotations for the Affymetrix whole transcript arrays, Omics Explorer will create simplified annotations (Gene Symbol, Gene Title,…) in addition to the annotations in the downloaded annotation file.

How to download and import annotations from the Affymetrix NetAffx server [6]

Select the **File → Download→ Affymetrix Probe Set Annotations** menu item.

Select the array type and press OK .

Select the annotations you want to import and press OK.

11. R USERS

Two R-scripts are shipped with Qlucore Omics Explorer. They make it easy to convert data back and forth from R.

The script function definition files are located in the user's Documents folder. The path is Documents/Qlucore/R-import/, where the two files read.gedata.1.1.R and write.gedata.1.1.R can be found.

For detailed description on how to use the scripts see the "R to .gedata" section in the Documentation and Help system.

12. TEMPLATES

Templates are scripts that enables a pre-configured set of operations to be executed in the program. Templates are based on Python.

Templates can be opened from **File → Execute Template.**

The Template Browser (**File → Template Browser**) shows an overview of all templates available at the location(s) specified in the Template section of the Preferences.

---

[6] The first time you use this function, you will also be asked to provide: your *Affymetrix user name* and your Thermo Fisher (*Affymetrix) password* and the folder where the downloaded annotation files will be stored. If you do not have an Affymetrix user name and password, you can get one by pressing the Register button. This will take you to the registration page on the Affymetrix web site.

13. NGS DATA

With the NGS module can a rich suite of different NGS related files be imported and pre-processed to build up a NGS project. Details are presented in the Documentation and Help system that is shipped with the program.

TERMINOLOGY

Samples: We use samples to describe units such as patients, persons, animals, treatments, dates,…

Variables: We use variables to describe quantities that have been measured for each samples, such as Gene Expression Levels, Protein concentrations, answers to a question in a questionnaire,…

Annotation: A description of a sample or a variable. One sample or variable can be described by one or many annotations.

DECIMAL POINT AND DECIMAL COMMA

Both separators are supported in all tab separated .txt files and .csv files.

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.

TRADEMARK LIST

Excel, Windows Vista, Windows 7 and Windows 10 are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

NetAffx is the trademark of Thermo Fisher Scientific (Affymetrix).

GenomeStudio and BeadStudio are trademarks of Illumina