How to analyze Proteomics data

This document introduces import and statistical analysis of proteomics data, as an example data from liquid chromatography - mass spectrometry (LC-MS), in Qlucore Omics Explorer (QOE). It is assumed that the user is familiar with the basic functionality of QOE.

In addition to the statistical tests that are highlighted below other functionality such as plots and visualizations are also relevant for proteomics data analysis.

Content

1. DATA PREPARATION

Proteomics data can be imported as un-normalized or normalized data. This document focuses on un-normalized data.

The Wizard can import tab separated .txt files, or comma separated .csv files. If your data is in a Microsoft .xlsx or .xls file you need to open the data and save it in a txt or csv format.

The normalization method supported for proteomics data is Quantile Normalization.

After completing the import through the Wizard, you select the Data tab to prepare your data before analysis begins.

2. IMPORT A PROTEOMICS DATA FILE

The program provides import of data and annotations through Wizards. Start QOE, select the file menu and then "Open with Wizard…" . The Wizard and steps are described in detail in the reference manual and in several videos that can be viewed at qlucore.com (Tip: Search for "Wizard" on https://qlucore.com/videos).

The data in the .csv file may be organized as vectors (columns) or as a matrix, the Wizard can handle both. After import, you save the imported data as a .gedata file.

3. THE METHODS TAB

After completing the import through the Wizard, you can select to show your data in a table view. This can give you a feeling for the data to be analyzed. To do so, select the Table plot in the Methods tab.

Then switch to the Data tab to prepare your data before analysis begins.

4. THE DATA TAB

4.1. OVERVIEW

The Data tab presents the data processing functionality. Data is processed in the same order as the tab is organized – from left to right.

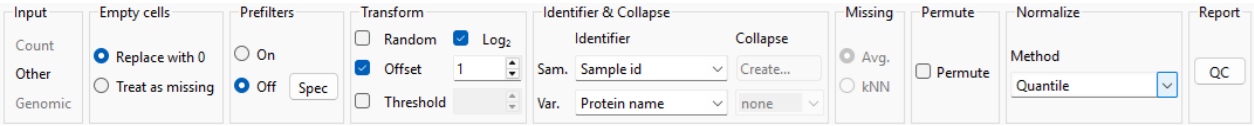

***Figure**: The Data tab*

4.2. EMPTY CELLS

Your data may contain cells that are empty. Such a cell may in your data be represented as empty, or as a "NA" or "N/A" value. These cells are in the Table plot indicated with a number in *Italic*.

An empty cell may indicate that the equipment detected a value below a detection limit, and the resulting cell was registered as empty. You can in this step select "Replace with 0", whereby the empty cells get a zero (0) value.

4.3. PREFILTERS

Prefilters allow you to take various actions on your data. You may not need to do any pre-filtering, but it can be useful in some situations. You can select to remove variables that are missing, and to remove variables with too high or too low values.

4.4. TRANSFORM

Proteomics data may after a selection of ”Replace with 0” contain a number of cells with zero (0) values.

As a part of the process of normalizing the data, the data is often log transformed. You can do this under Transform.

Zero values cannot be log transformed, so before log transformation, the values need to be larger than 0. There are two ways to achieve that:

- Offset. Adds an offset value (like 0.01 or 1) to every cell in the dataset. This will uplift the dataset values and make log transformation possible. The added value must be high enough to lift every cell value above 0.

- Threshold. Here is set the minimum value of every cell to a fixed value (like 0.01 or 1). Only those cells with a value lower than the Threshold value are affected.

The advantage with the Offset approach compared with the Threshold approach is that distances between values are kept after transformation.

To log transform your data, select the checkbox “Log_2”.

4.5. IDENTIFIER & COLLAPSE

Data may have variables that are not unique. In this step, you make your variables unique by selecting a method to calculate the collapsed value (using max, min, average, median or sum) of the variables that are not unique.

The variable identifier can be changed, in which case a collapse may be a logical step thereafter.

Sample collapse is also supported. As an example, if you have replicates from the same patient, you may change the sample identifier from “Sample” to “Patient” and collapse on the sample side (max, min, average or median). The result is saved as a new data set.

4.6. MISSING

Selection of method for missing value reconstruction, values are imputed. The option is not active if the “Empty cells” option has been used.

4.7. NORMALIZE

A traditional method for normalization of Proteomics data is Quantile Normalization. The method makes the distributions of the variables across the samples become the same and they can therefore be compared.

4.8. EXAMINE THE TABLE VIEW

In Table plot the actual values as a result of all data processing is presented.

4.9. SAVE THE DATASET

Remember to save the dataset after your final changes. Select “Save as …” under “File” in the top menu of the Program. If you use the same file name as before you will overwrite the originally imported dataset.

5. ADJUSTMENT OF SAMPLES

Go to the Samples tab to examine samples and sample annotations.

The dataset may have control samples included in the same run as your own samples. These control samples can be excluded from the analysis. You can then uncheck these samples in the Samples tab.

6. ADD MORE ANNOTATIONS TO YOUR SAMPLES

You can import additional sample annotations. To add the new annotation information, go to “File” in the top menu, select “Import” and then “Sample Annotation via Wizard …”.

7. DATA OVERVIEW IN PLOTS

A heatmap can often give a good overview of the data. Change the plot to a heatmap in the Methods Tab. Different ordering options are available both on the sample and the variable side.

For more information about individual variables use for instance Bar and Violin plots.

8. STATISTICAL TESTS

8.1. A TWO GROUP TEST

A two-group comparison can be used to test the significance of the difference between two conditions. QOE supports several two group tests: t-test, Welch test and Mann Whitney U Test (Wilcoxon Rank Sum Test)

The standard t-test assumes that the input values are normally distributed and the variance between the groups is comparable. The Welch test is an alternative when one group has many more samples than the other. The Mann Whitney U Test (Wilcoxon Rank Sum Test) is used when data is not normally distributed,

To eliminate the influence from covariates, use the eliminated factor option and select the sample annotation to eliminate.

8.2. A REGRESSION TEST

For numerical or ranked annotations there are several types of regression tests supported.

9. ACKNOWLEDGEMENTS

Excel is a trademark of Microsoft Corporation.

10. 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.