

# **Tutorial for proteome data analysis using the Perseus software platform**

Laboratory of Mass Spectrometry, LNBio, CNPEM

Tutorial version 1.0, January 2014.

Note: This tutorial was written based on the information available in scientific papers, MaxQuant google groups, local group discussions and it includes our own experiences in the proteomics data analysis performed in our research group.

## Release information

Perseus Tutorial version 1.0, January 2014.

Software versions discussed in this tutorial:

MaxQuant version 1.3.0.5 from 2012.

Perseus version 1.3.0.4 from 2012.

Software updates can be obtained through the MaxQuant website:

<http://www.maxquant.org/>

## Contents

Release information.....	2
1. Perseus.....	4
1.1. Getting started with Perseus .....	4
1.2. Installing Perseus software .....	4
1.3. Loading MaxQuant results data into Perseus.....	4
1.4. Data frame selection .....	5
1.5. Data pre-processing .....	8
1.5.1. Data filtering.....	8
1.5.2. Data transformation .....	9
1.5.3. Categorical annotation of rows.....	10
1.5.4. Filtering valid values .....	12
1.6. Data correlation and quality check .....	13
1.7. Principal component analysis .....	14
1.8. Statistical analysis.....	15
1.8.1. One-sample T- test.....	15
1.6.2. Two-sample T-test.....	16
1.6.3. Multiple samples tests .....	18
1.7. Data visualization .....	18
1.7.1. Numeric Venn diagram .....	18
1.7.2. Building heat maps .....	19
1.8. Saving your analysis and results .....	22

## 1. Perseus

### 1.1. Getting started with Perseus

Perseus is a software framework for the data annotation and statistical analysis of proteomics data obtained through high-resolution Mass Spectrometry. It was written in C# using the .NET Framework 4.5. It runs in Windows operational systems (Windows 7 or higher) and Windows Vista SP2. It can also run in Windows server 2008 or 2012. NET framework 4.5 must be installed in your computer before installing the Perseus software. The program can execute four main types of activities: Uploading, Processing, Analysis, Combination and Exporting. Perseus is a program developed as a plugin; therefore, you may also contribute to develop the tools through the project web page (<http://jurgencox.github.io/perseus-plugins/>).

### 1.2. Installing Perseus software

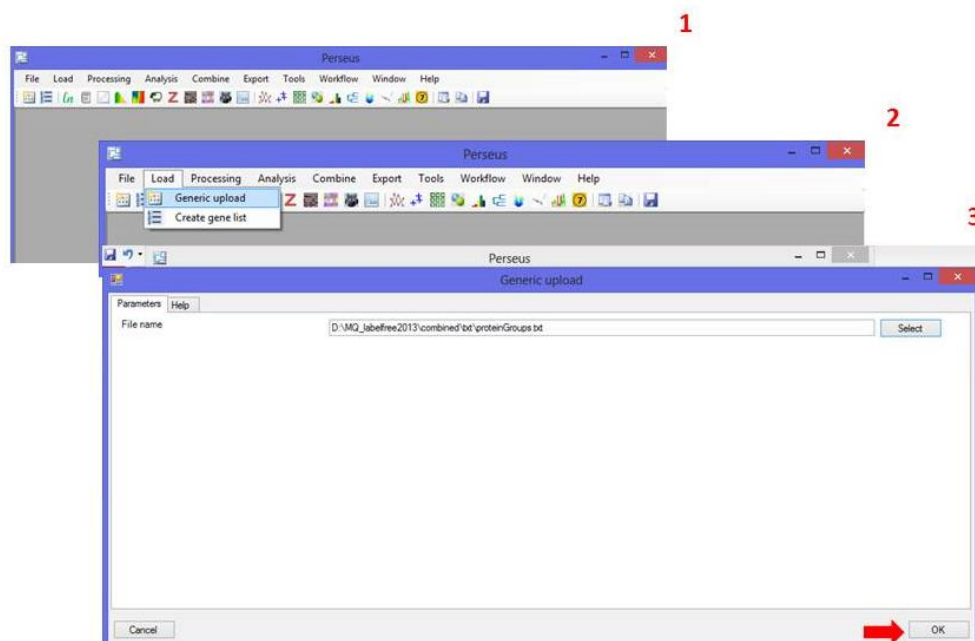
1. Go to the web site of Perseus (<http://www.perseus-framework.org/>)
2. You have to register in the website to have access to the most recent version of Perseus or you can contact us directly for a copy of the version of Perseus shown in this tutorial (v.1.3.0.4).
3. You will be asked to fill a registration form and you will receive by e-mail a code for downloading the Perseus software.
4. After downloading is complete, unzip the files in the new created folder (e.g., Perseus\_1.3.0.4).

### 1.3. Loading MaxQuant results data into Perseus

1. Double click on the Perseus.exe file to execute the program;
2. In the main menu bar, open the **Load** menu and select the Generic upload;
3. Click on the **Select** button and browse the folder where your MaxQuant results were saved. Go to the folder **combined** > **txt** and select the data file "proteinGroups.txt" and click OK.

A new window should open containing five feature tabs (Expression, Categorical annotation, Textual annotation, Numerical annotation, Multi-numerical annotation), where the information you want to include in the data analysis will be chosen by you.

**Attention:** In this particular example we are going to perform a Label-Free quantification analysis of proteomics data. No protein labeling was applied to the proteins analyzed in this study. All analysis further demonstrated here in this tutorial are focused in the Label-Free proteomic analysis.

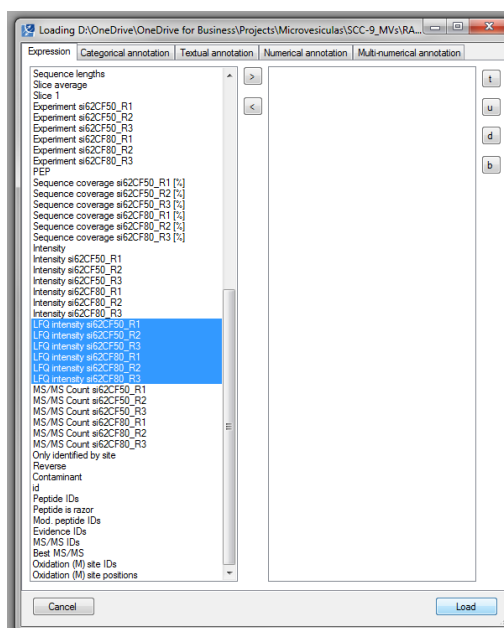


#### 1.4. Data frame selection

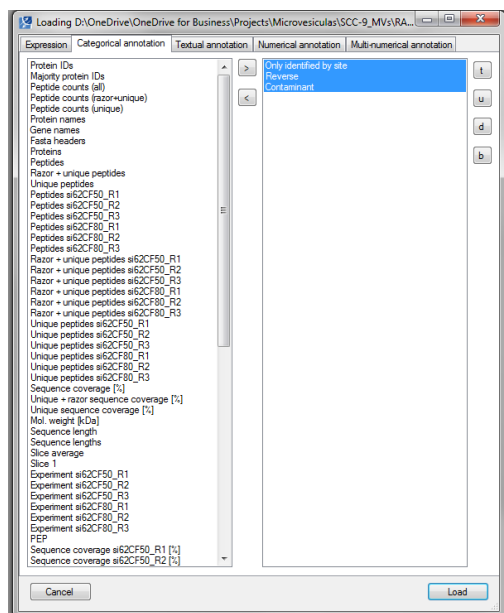
The Perseus platform was design to downstream bioinformatics, as well as statistics analysis on pre-processed data. Therefore, the data features must be selected carefully to compose the data frame which will be processed.

Once the data file is loaded, the data features must be selected (**Figure 2**).

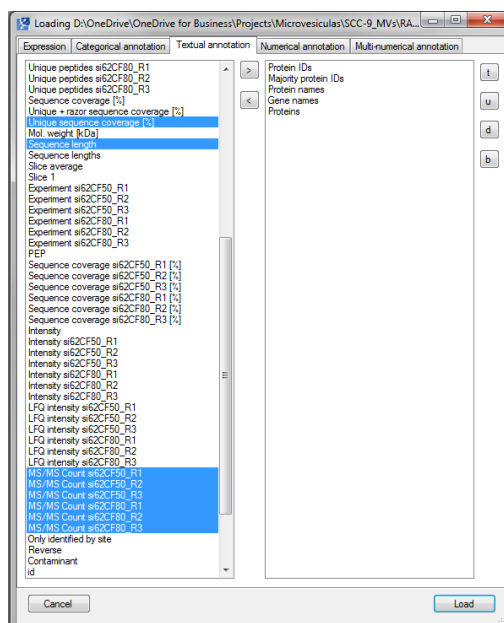
1. In the Expression tab, select values starting with **LFQ intensity** in the left box and click in the symbol “>” to transfer it to the right box. Remember that Expression values are the ones which will be expressing the protein abundance in a quantitative proteome analysis, and will be used in further statistical analysis.



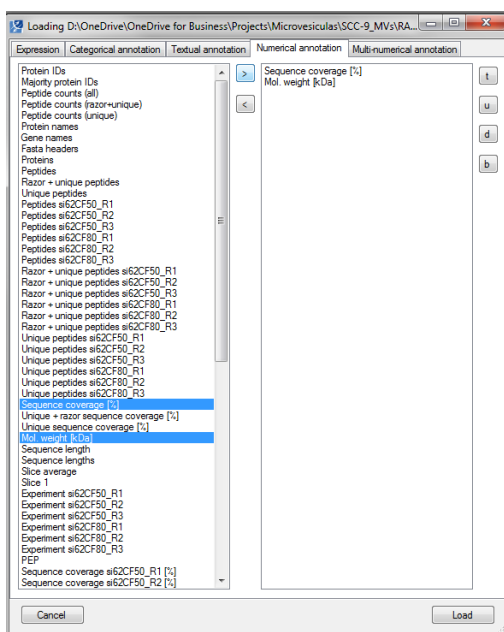
2. In the Categorical annotation tab, you will see three default parameters (Only identified by site, Reverse and Contaminant) which refer to identified hits which can be further filtered out of the data.



3. In the Textual annotation tab, five features are included by default (Protein IDs, Majority protein IDs, Protein names, Gene names, Proteins). Other features such as other expression values, unique sequence coverage and sequence length might be included in the data frame as text annotation.



4. In the Numerical annotation tab, select Mol. Weight (kDa) and Sequence coverage (%) and click in the symbol “>” to transfer the features from the left to the right box.



5. In the Multi-numerical menu, leave the right box in blank.
6. Click in **Load**.
7. The protein identification data will appear containing all data selected during the previous steps shown above within a data matrix (Matrix1), in the **Data** tab. Each step of the data analysis will generate a new data matrix. So, remember

that all data matrixes are available for you. To have access to these matrixes you just have to minimize or maximize the data matrixes you need.

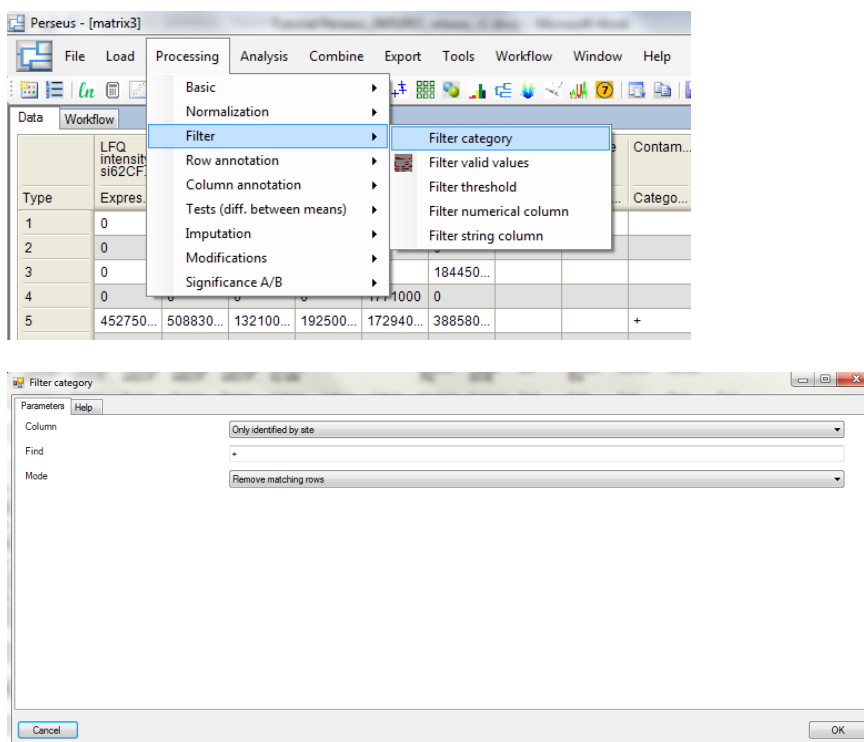
Another tab named **Workflow** appears and can be used to track back all actions taken during the data pre-processing and analysis. The workflow can be further saved by clicking in the menu **Workflow > Save As...**

## 1.5. Data pre-processing

### 1.5.1. Data filtering

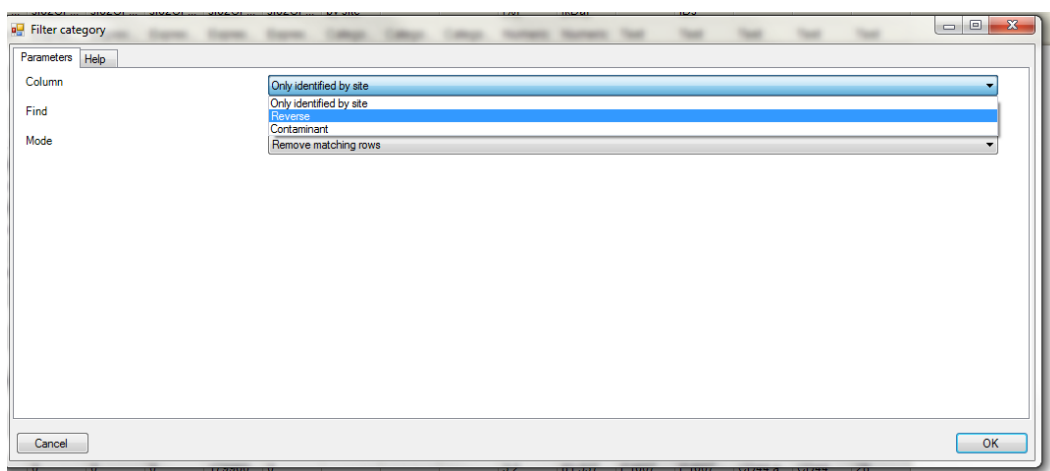
In this step the unnecessary or incorrect protein identifications can be removed from the main data frame. The protein identifications classified as **Only identified by site**, **Contaminants**<sup>1</sup> and **Reverse** can be excluded from the data frame.

1. In the Data matrix go to **Processing > Filter > Filter category**. In the *Column* parameter select **Only identified by site** and verify that *Find* parameter contains the symbol “+” and the *Mode* parameter contains **Remove matching rows**. Click OK.



<sup>1</sup> **Note:** Depending on the experimental design some of these categories could be of interest to keep in the final data frame, such as the protein hits identified as Contaminant proteins.

2. In the Data matrix go to **Processing > Filter > Filter category**. In the *Column* parameter select **Reverse** and verify that *Find* parameter contains the symbol “+” and the *Mode* parameter contains **Remove matching rows**. Click OK.

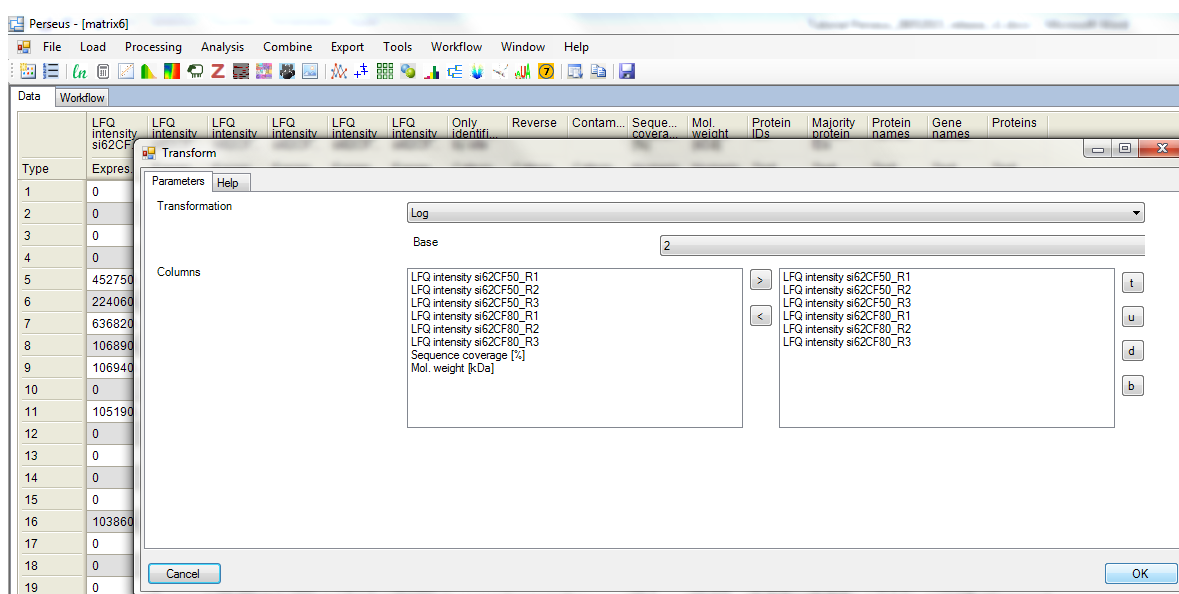
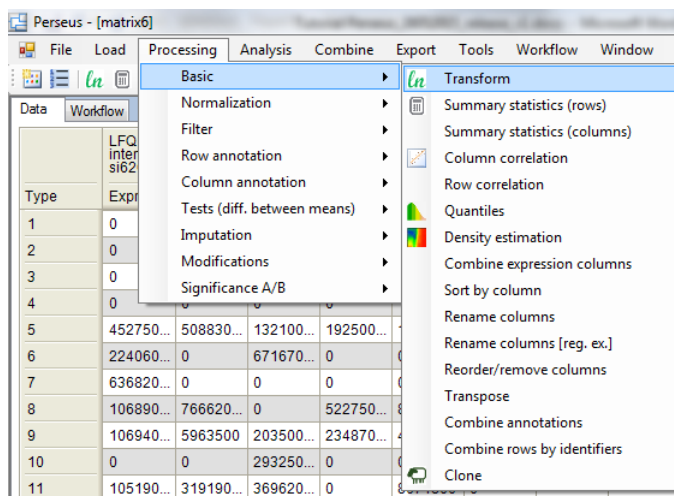


3. If you also wish to exclude all proteins assigned as Contaminants, you must go to **Processing > Filter > Filter category**. In the *Column* parameter select **Contaminant** and verify that *Find* parameter contains the symbol “+” and the *Mode* parameter contains **Remove matching rows**. Click OK.

### 1.5.2. Data transformation

As the range of the expression values can vary more than 10 folds, the expression values can be Log transformed in order to facilitate the calculation of the protein expression fold change.

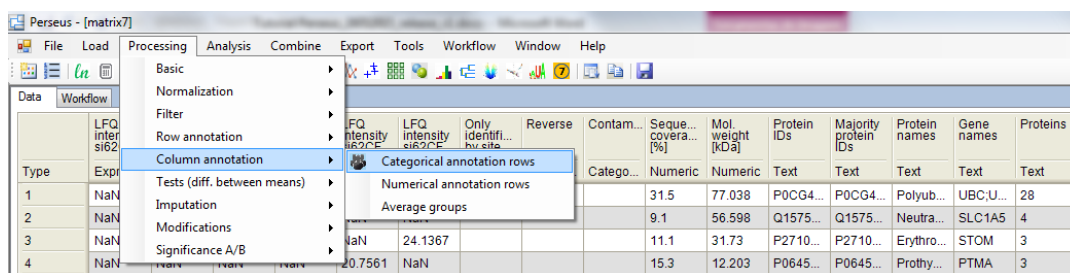
1. Go to **Processing > Basic > Transform**. In *Transformation* parameter, select **Log** and in the *Base* parameter select **2**. In the *Columns* box, select to which expression values the transformation should be applied and click on the symbol “>” to transfer the features from the left to the right box. Click OK.



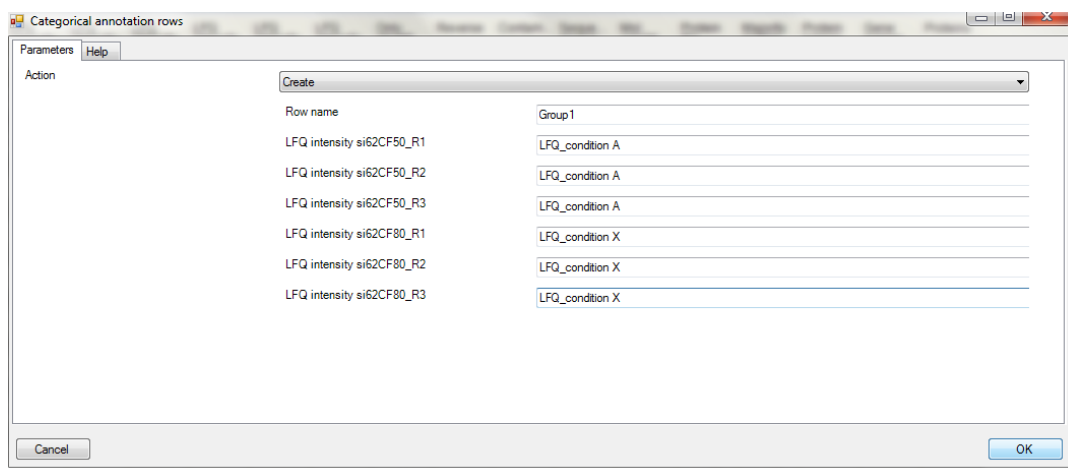
### 1.5.3. Categorical annotation of rows

In order to group the conditions compared in this quantitative analysis you must annotate the categories/classes of the conditions analyzed into groups, which will be further compared.

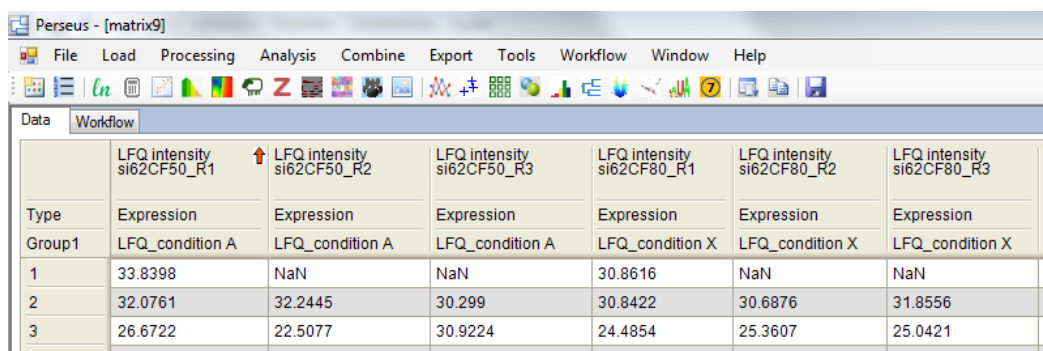
1. Go to **Processing > Column annotation > Categorical annotation rows**.



2. In the *Action* parameter, select **Create** and include one name for the grouping you will perform. By default the Row name is **Group 1**.
3. Below the Row name, the names of the expression features will appear and they must be substituted by the names which will define the conditions/classes to which the expression value belongs. In this way, expression values of biological replicates from one condition must have the same group name (e.g., Labeled condition A), and the condition against it is been compared must all have the same group name (e.g., Labeled condition X).



4. After naming the groups, click OK.
5. Make sure that the replicates belonging to the same condition received the same name in the general data frame. This step defines the groups which will be analyzed by statistical tools. After this selection, the data frame with grouping categories should appear as shown below.

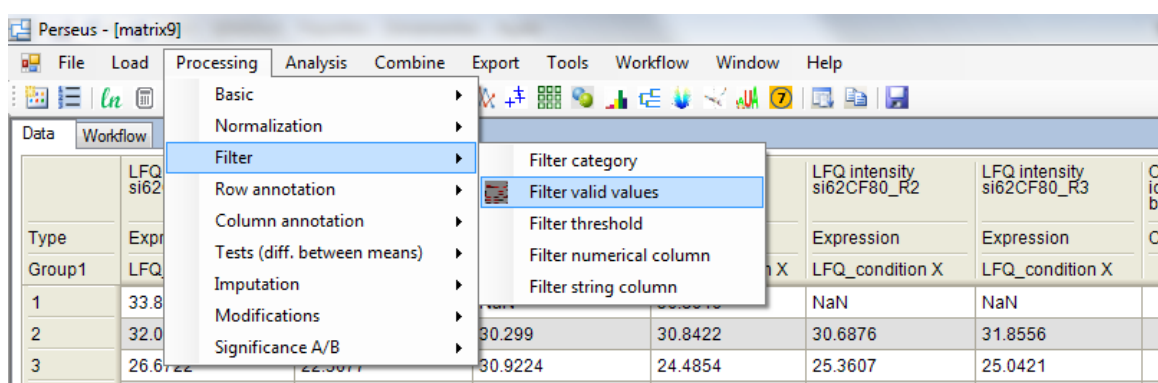


	LFQ intensity si62CF50_R1	LFQ intensity si62CF50_R2	LFQ intensity si62CF50_R3	LFQ intensity si62CF80_R1	LFQ intensity si62CF80_R2	LFQ intensity si62CF80_R3	
Type	Expression	Expression	Expression	Expression	Expression	Expression	
Group1	LFQ_condition A	LFQ_condition A	LFQ_condition A	LFQ_condition X	LFQ_condition X	LFQ_condition X	
1	33.8398	NaN	NaN	30.8616	NaN	NaN	
2	32.0761	32.2445	30.299	30.8422	30.6876	31.8556	
3	26.6722	22.5077	30.9224	24.4854	25.3607	25.0421	

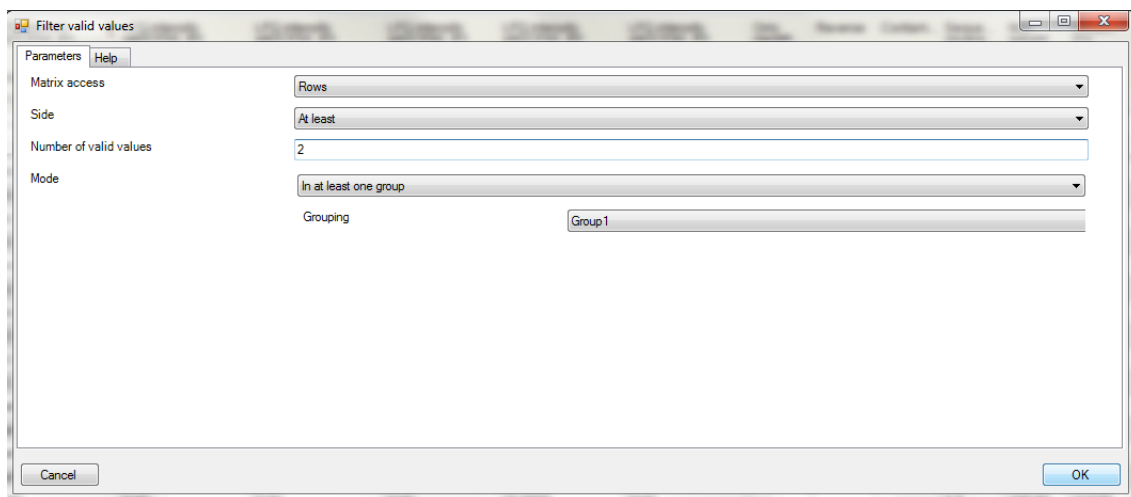
#### 1.5.4. Filtering valid values

The Log transformation of the expression values generate a pool of “NaN” (Non-Assigned Number) values, which correspond to expression values originally equal to zero, when proteins were not detected by the mass spectrometer. In order to define the level of stringency of your analysis, you must define the minimum number of valid values accepted in your analysis. This may increase the confidence of your data. For example: If you want to compare two biological conditions such as control versus treatment, and you have performed 3 biological replicates for each condition, you must define in how many biological replicates you have to identify the same protein to consider that this protein on further statistics or annotation processes. As more hits you request for certain protein identification, the more stringent is your criteria. Nevertheless, remember that in any case a protein might be present in only one biological condition. For example, some proteins may only be present in the group of control condition and totally absent in the treatment group.

6. Go to **Processing > Filter > Filter valid values**.



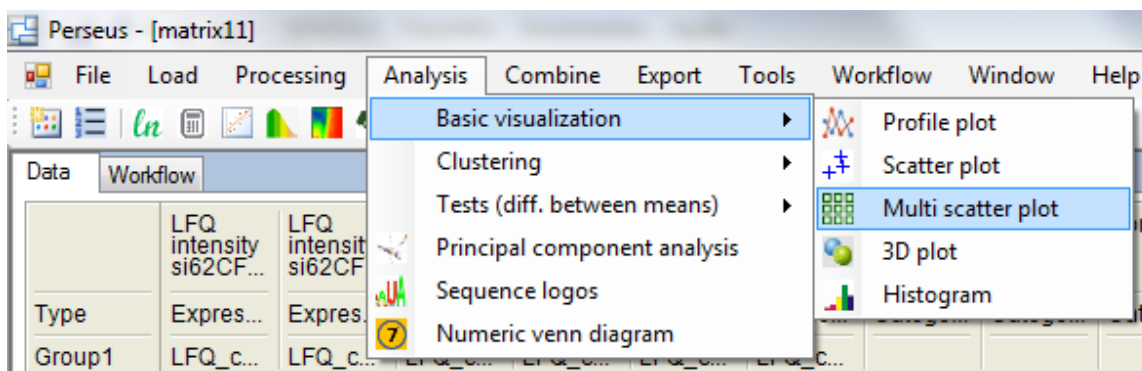
7. In the *Matrix access* parameter, select **Rows** and in the *Side* parameter select **At least**. In the *Number of valid values* parameter type the number **2**. In the *Mode* parameter select **In at least one group**.



### 1.6. Data correlation and quality check

Once you have transformed the LFQ intensity values to Log2 values and valid values were filtered, you can estimate the degree of correlation between the different samples and to each extend the replicates are similar or dissimilar to each other. This analysis may help you to eliminate datasets that behave as outliers, since the values of data correlation will be calculated and help you to identify these possible outliers.

1. Go to **Analysis > Basic visualization > Multi scatter plot**.
2. Select to which conditions and replicates the Multi scatter plot should be generated, by including or excluding replicates from the right side box.
3. Press OK.



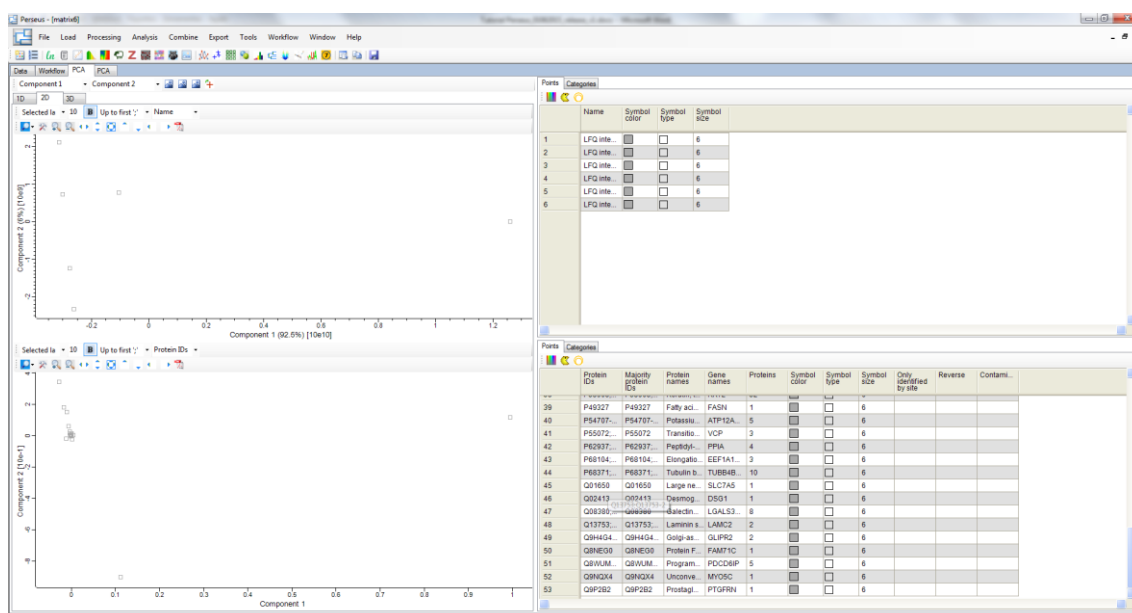
A Multi scatter plot must appear in a new Tab of the main data analysis window. In the Multi scatter plot you can observe that small numbers are indicated inside each individual scatter plot (e.g.,  $r=0.891$ ), which indicates the calculated correlation between the datasets indicated in the y and x axis. As closer to 1, as higher is the correlation between the two datasets compared.

Single scatter plots can also be generated through the selection of the **Analysis > Basic visualization > Scatter plot**.

### 1.7. Principal component analysis

One of the strategies useful to identify variations between different conditions is to detect similarities or dissimilarities among the expression patterns of the proteins using Principal Component Analysis.

1. Open the data matrix where you have the data filtered for the main pre-processing steps for data clean up and the LFQ intensity values are **not** transformed to Log2. This may be in one of the initial data matrixes you have in your data analysis.
2. Go to **Analysis > Principal component analysis**.



In a new Tab, the PCA results will be shown and you can observe the participation of the principal components in explaining the data variance in terms of percentage, shown

the y and x axis. The lists of the variables (proteins) which contribute to the variations observed in the data are listed in the right side boxes.

### **1.8. Statistical analysis**

For determining the variations between the proteome from two or more conditions, it is necessary not only to identify the proteins present in a sample, but also to perform statistical tests to determine if the changes observed experimentally in the expression values of the proteins are statistically significant.

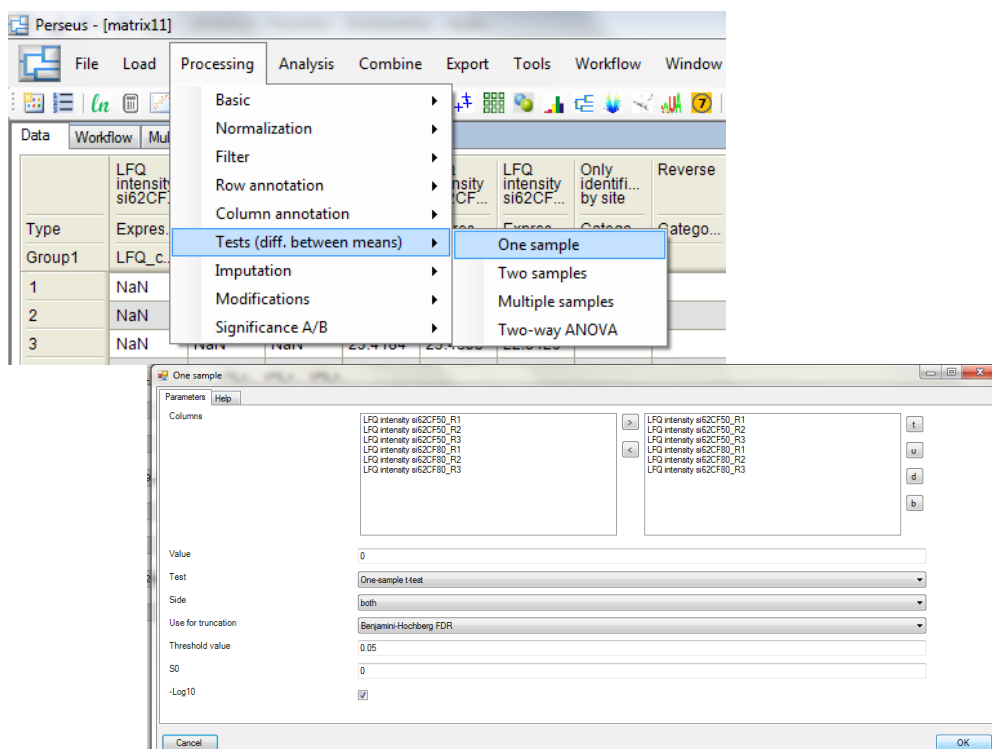
Perseus offers you at least four different statistical tests that can be performed for the proteome data analysis.

#### **1.8.1. One-sample T- test**

One-sample t-test compares the mean score of a sample to a known value; usually the populations mean, in other words, the one-sample t-test, in which the level of outcome for a group is compared to a known standard. The basic idea of the test is a comparison of the average of the sample (observed average) and the population (expected average), with an adjustment for the number of cases in the sample and the standard deviation of the average.

In the Perseus, One sample-test will define if the mean of the LFQ intensity values measured is significantly different from a fixed value (typically 0). After this process, two numerical columns will result in the new data matrix, one containing the T-test p-value (-Log t-test p-value), the other one containing the difference between the mean and the fixed value. In addition, there is a categorical column added where the symbol '+' appears when the variation between the groups is statistically significant with respect to the specified threshold.

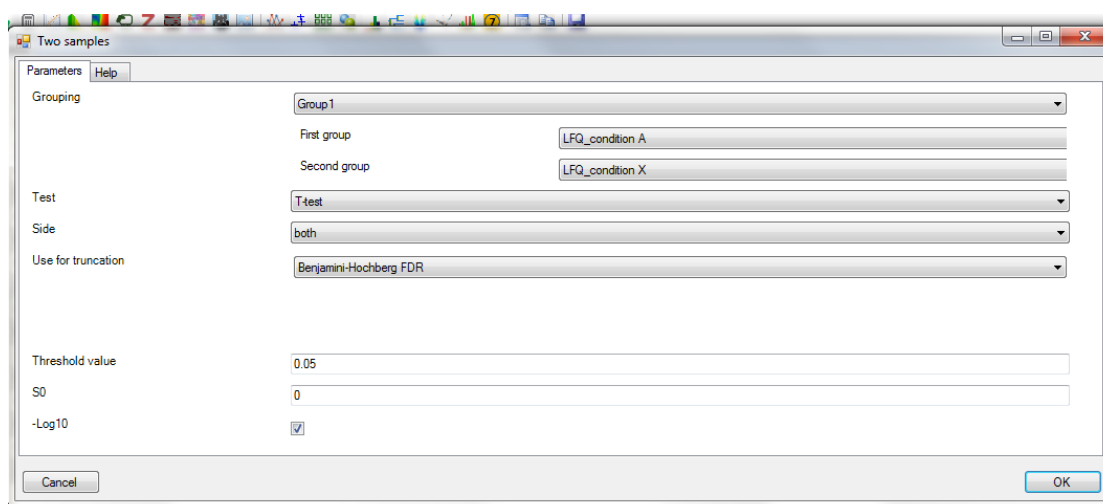
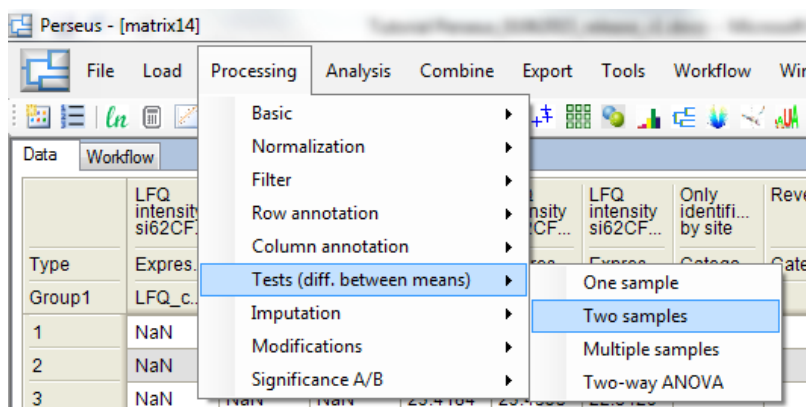
1. Go to **Processing > Tests (diff. between means) > One sample**.
2. In the **Columns** boxes, select to which expression values the test must be applied by including or removing expression columns in the right side box.
3. In the **Value** cell, keep the default **zero**.
4. In the **Test** cell, keep the **One-sample t-test** default option.
5. In the **Side** cell, keep the default **both** option.
6. In the **Use for truncation** cell, select **Benjamini-Hochberg FDR**.
7. In the **Threshold value** include the **0.05** (5% probability error).
8. Keep the **S0** cell with default value **zero**.
9. Click OK.



	LQ intensity si62CF...	LQ intensity si62CF...	LQ intensity si62CF...	LQ intensity si62CF...	LQ intensity si62CF...	LQ intensity si62CF...	Only identi... by site	Reverse	Contam...	t-test Signific...	Seque... covera... [%]	Mol. weight [kDa]	-Log t-test p value	t-test Differe...
Type	Expres...	Expres...	Expres...	Expres...	Expres...	Expres...	Catego...	Catego...	Catego...	Catego...	Numeric	Numeric	Numeric	Numeric
Group1	LQ_c...	LQ_c...	LQ_c...	LQ_c...	LQ_c...	LQ_c...								
1	NaN	NaN	23.6338	26.7376	26.8284	25.6405				+	22.4	28.036	4.27539	25.7101
2	NaN	NaN	NaN	24.2267	25.4615	NaN				+	1.7	272.32	1.80085	24.8441

### 1.6.2. Two-sample T-test

Two-sample test is applied for determining if the means of the LQ intensity values of two categories/classes of samples are significantly different from each other. As a result, two numerical columns are further added in the data matrix, one containing the p-value, the other one containing the difference between the means. Again, there is a categorical column added containing a symbol '+' when the changes in protein abundance between the different groups is statistically significant with respect to the specified threshold.



Methods for calculating statistical confidence are available and can be applied for the proteomics data. Two of these methods include multiple testing corrections.

### P value:

The P value or calculated probability is the estimated probability of rejecting the **null hypothesis ( $H_0$ )** when that hypothesis is true. The null hypothesis is usually the result of "no difference" between two groups. If the P-value is less than the chosen significance level then the null hypothesis is rejected, i.e., the **alternative hypothesis** is accepted suggesting that there is a difference between the mean values from treatment and control data.

### Permutation-based FDR:

Permutation based procedure, is used to adjust for multiplicity tests by controlling the family-wise type I error rate (FWER) - which is the probability of making one or more

false discoveries, or type I errors among all the hypotheses when performing multiple hypothesis tests - without assuming  $t$  distribution of the test statistics of each gene's differential expression.

#### **Benjamini-Hochberg FDR:**

Calculates the False Discovery Rate (FDR) for each p-value generated through independent tests. However, this statistical method is applicable in multiple comparisons, in other words, in multiple hypothesis test that a null hypothesis could be rejected incorrectly when considering a data set as a whole.

### **1.6.3. Multiple samples tests**

Statistical analysis for multiple biological conditions can also be performed using Multiple samples t-test and One-way or Two-way ANOVA, which are also available through the same path shown for the t-tests (**Processing > Tests (diff. between means)**)).

## **1.7. Data visualization**

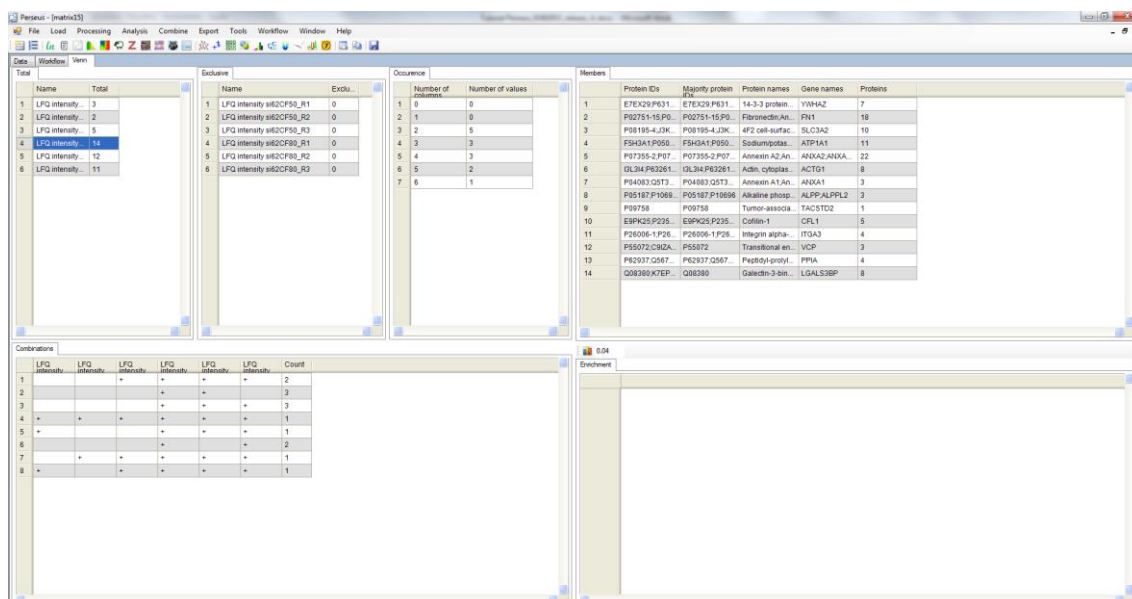
Once the differentially expressed proteins are confirmed with statistical methods, other tools can be applied for the analysis of the biological meaning of the results. Proteins exclusively identified in one condition, proteins differentially expressed or proteins present in certain combinatorial conditions can be selected for further analysis.

### **1.7.1. Numeric Venn diagram**

Once the categories and grouping of conditions have been assigned, the combinations and number of proteins identified on each group or condition can be calculated based on the information provided by numeric Venn diagram function.

1. Go to **Analysis > Numeric venn diagram**.

A new window should appear.



In the **Total** tab, the number of identifications per replicate is listed and the protein identifications belonging to these lists can be seen in the **Members** tab. In the **Exclusion** tab, the numbers of identifications exclusive to each replicate and condition are listed and can be observed in the **Members** tab, once selected.

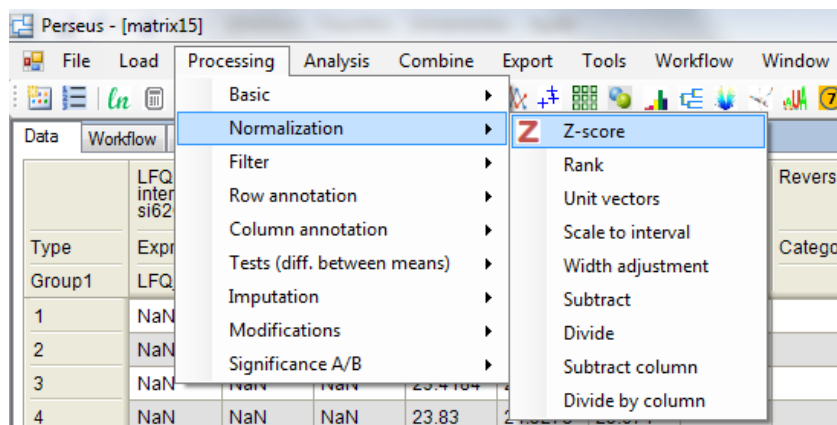
Identifications reported in the Members tab can be totally or partially selected using Ctrl key and exported by right clicking the mouse button over the table.

This data can then be used to create Venn diagrams to show the distribution of the total number of common or exclusive proteins identified on the different datasets.

### 1.7.2. Building heat maps

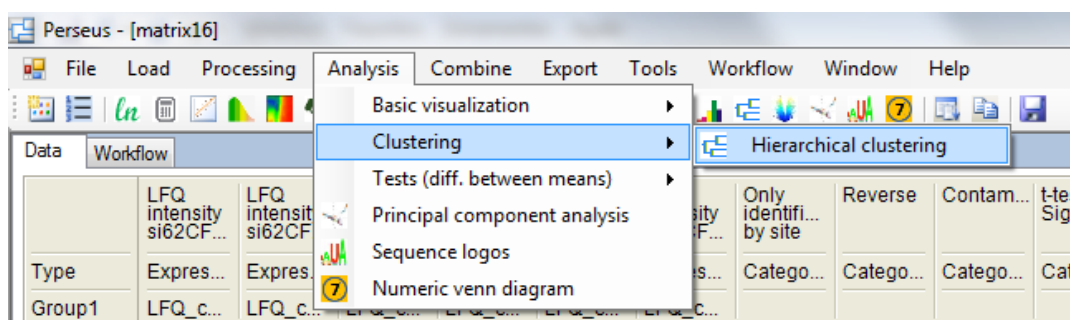
A heat map is a visual representation of the relative expression levels of the proteins according to a clustering behavior. In other terms, a heat map is a graphical representation of the data where the individual values contained in a data matrix are represented as colors. In order to graph the difference in proteins abundance we perform a series of actions for data clustering and visualization.

1. Go to **Processing > Normalization > Z-score**.
2. In the Parameters tab, select Rows.
3. Click OK.

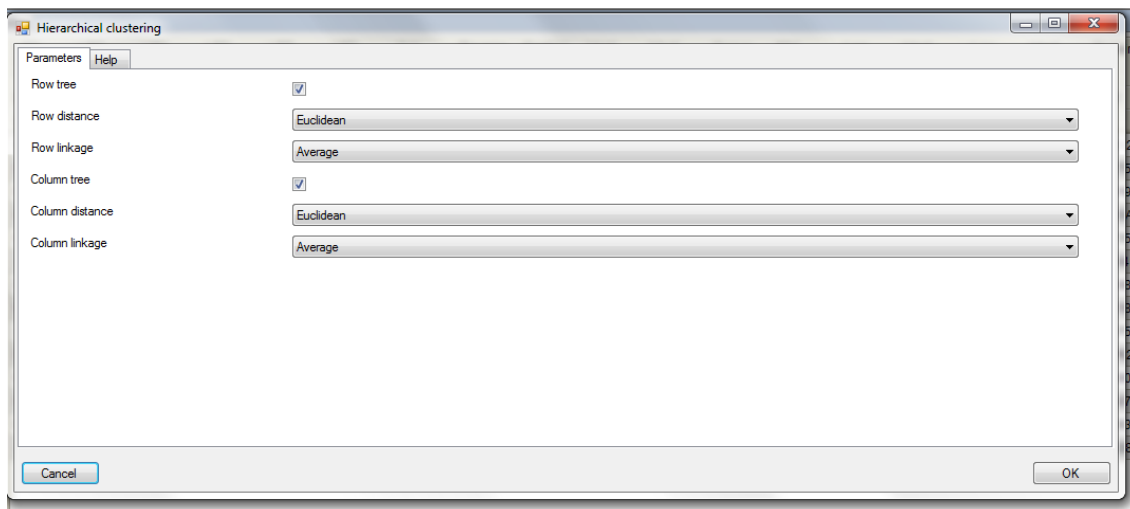


The expression data from each Row will be normalized by the mean value calculated for the same Row.

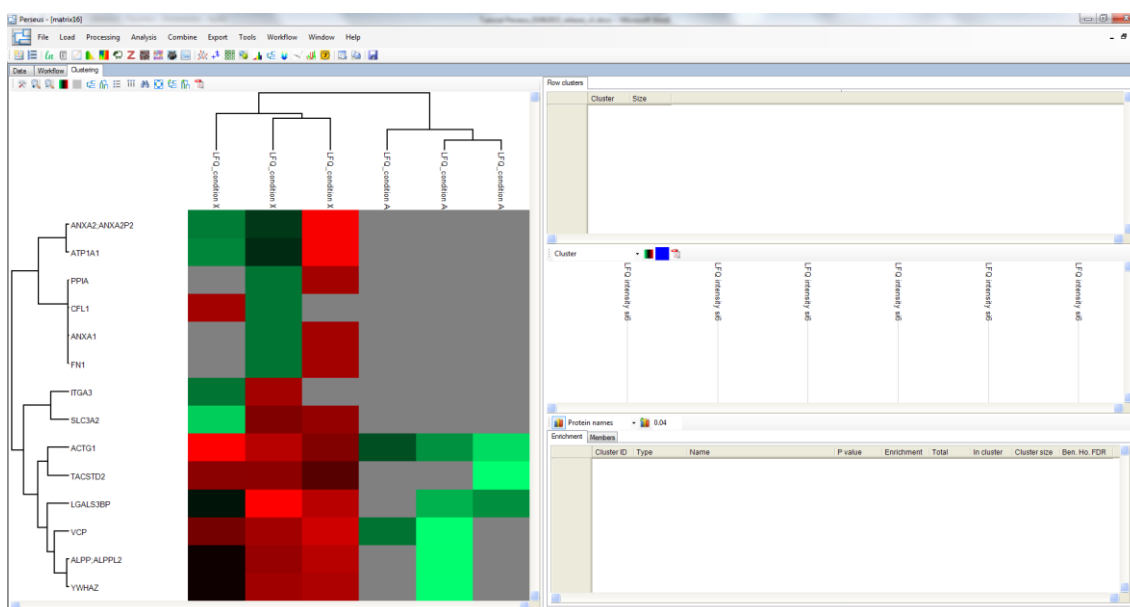
4. For creating the Heat map go to **Analysis > Clustering > Hierarchical clustering**.



5. Select the option **Row** tree.
6. In the Row distance cell, select **Euclidean**.
7. In the Row linkage cell select **Average**.
8. Select the option Column tree
9. In the Column distance cell, select **Euclidean**.
10. In the Column linkage cell, select **Average**.
11. Click OK.



A Heat map will be created based on the data clustering using Euclidean distance method in a new tab named **Clustering**.



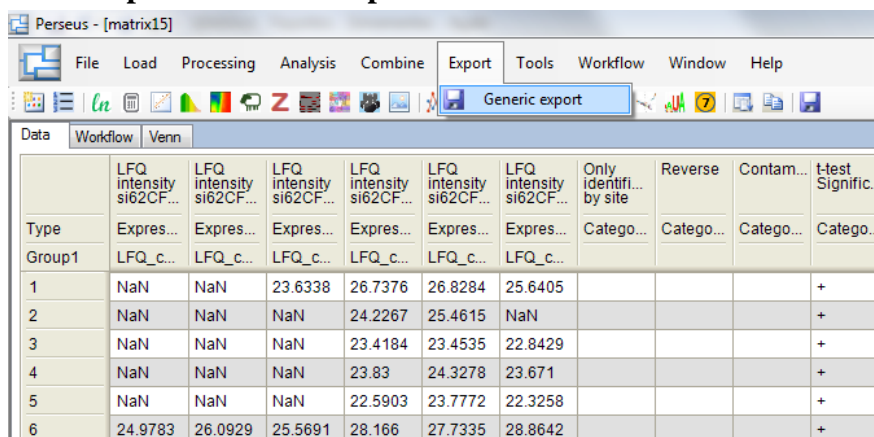
For modifying the names and other styling parameters of the Heat map you can modify the default settings in the tools available in the Clustering tab.

The Heat map image can be exported as a .pdf file by clicking in the small icon over the Heat map in the buttons of the parameter settings bar of the Clustering tab.

## 1.8. Saving your analysis and results

Once your analysis is complete or you want to finish the analysis later on, you can save any matrix you have by exporting it to a file.

### 1. Go to **Export > General export**.



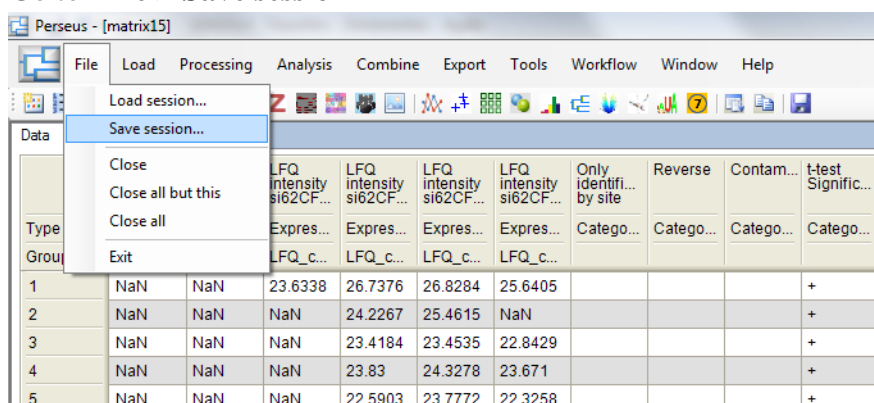
Perseus - [matrix15]

File Load Processing Analysis Combine Export Tools Workflow Window Help

Generic export

	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	Only identi... by site	Reverse	Contam...	t-test Signific...
Type	Expres...	Expres...	Expres...	Expres...	Expres...	Expres...	Catego...	Catego...	Catego...	Catego...
Group1	LFQ_c...	LFQ_c...	LFQ_c...	LFQ_c...	LFQ_c...	LFQ_c...				
1	NaN	NaN	23.6338	26.7376	26.8284	25.6405				+
2	NaN	NaN	NaN	24.2267	25.4615	NaN				+
3	NaN	NaN	NaN	23.4184	23.4535	22.8429				+
4	NaN	NaN	NaN	23.83	24.3278	23.671				+
5	NaN	NaN	NaN	22.5903	23.7772	22.3258				+
6	24.9783	26.0929	25.5691	28.166	27.7335	28.8642				+

- The results are saved as .txt files and can be either opened again in the Perseus software or in programs such Microsoft Excel.
- For saving the whole analysis, the Perseus session can be saved, also at any time of the analysis.
- Go to **File > Save session**



Perseus - [matrix15]

File Load Processing Analysis Combine Export Tools Workflow Window Help

Load session...  
Save session...  
Close  
Close all but this  
Close all  
Exit

	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	LFQ intensity si62CF...	Only identi... by site	Reverse	Contam...	t-test Signific...
Type	Expres...	Expres...	Expres...	Expres...	Catego...	Catego...	Catego...	Catego...
Group1	LFQ_c...	LFQ_c...	LFQ_c...	LFQ_c...				
1	NaN	NaN	23.6338	26.7376	26.8284	25.6405		+
2	NaN	NaN	NaN	24.2267	25.4615	NaN		+
3	NaN	NaN	NaN	23.4184	23.4535	22.8429		+
4	NaN	NaN	NaN	23.83	24.3278	23.671		+
5	NaN	NaN	NaN	22.5903	23.7772	22.3258		+

- For saving the sequence of data analysis you have performed and to apply it again, you can go to **Workflow > Save As..**

We recommend you to always save your first data matrix, multiscatter plot, heat map and your data matrix containing the results of the statistical tests for reporting purposes.

Your final data as .txt file can now be opened in other programs for further functional annotation of the candidate proteins.