



# An introduction to Perseus:

## Functional Enrichment Analysis

Goh Hoe Han, PhD

# Learning Objectives

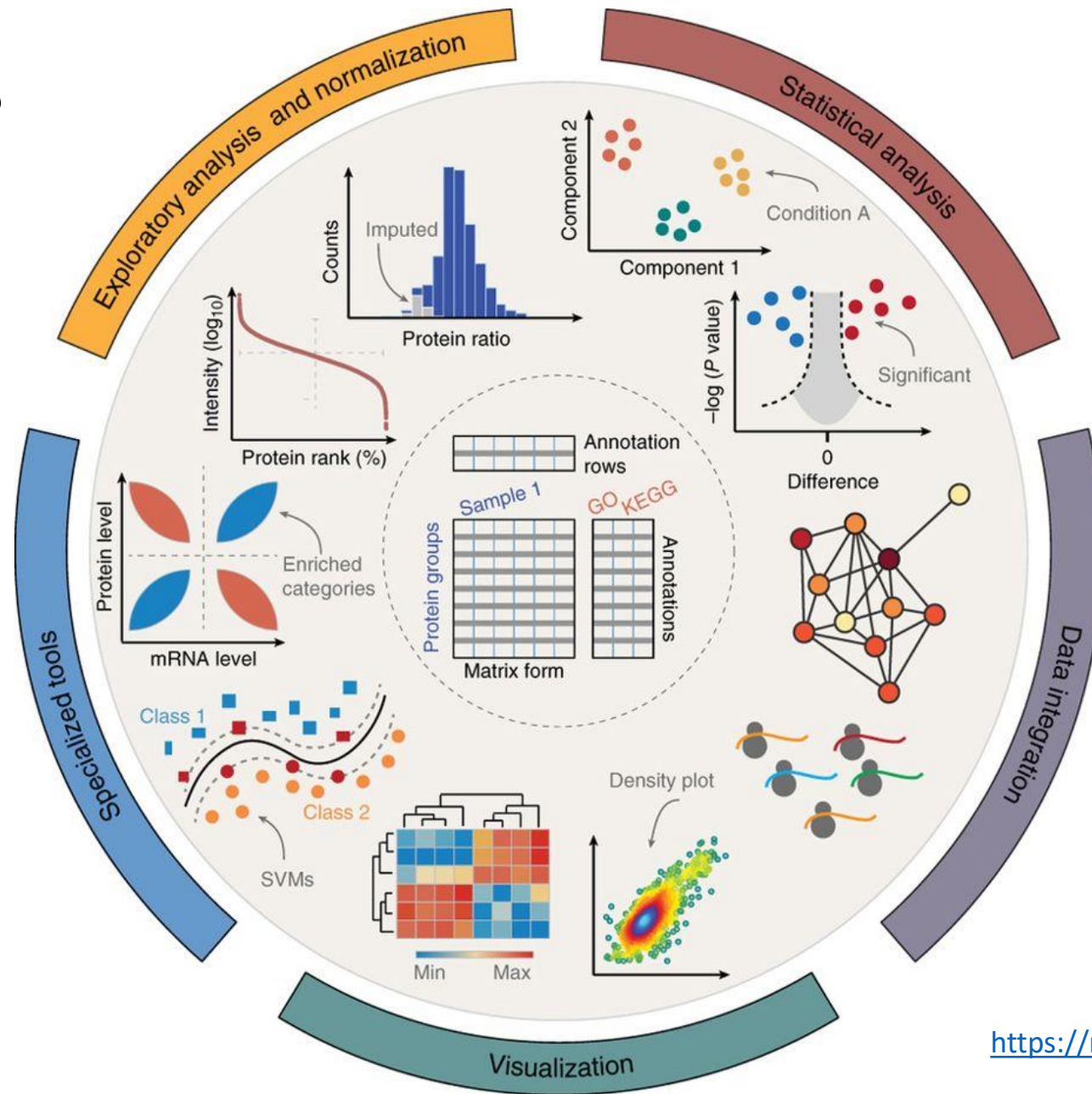
1. To **apply** knowledge of **functional enrichment analysis** through **Perseus** data analysis
  - A. To be familiar with the **basic functionality** of Perseus
  - B. To be able to perform functional enrichment analysis

# Outline

A stylized, light gray graphic of a DNA double helix is positioned on the left side of the slide, extending from the top to the bottom. It is partially obscured by the 'Outline' title.

- Background
  - Software - Perseus
- Demo
  - Loading the data
  - Filtering
  - Exploratory analysis
  - Loading annotations
  - Differential expression analysis
  - Clustering & Profile plots
  - Functional analysis

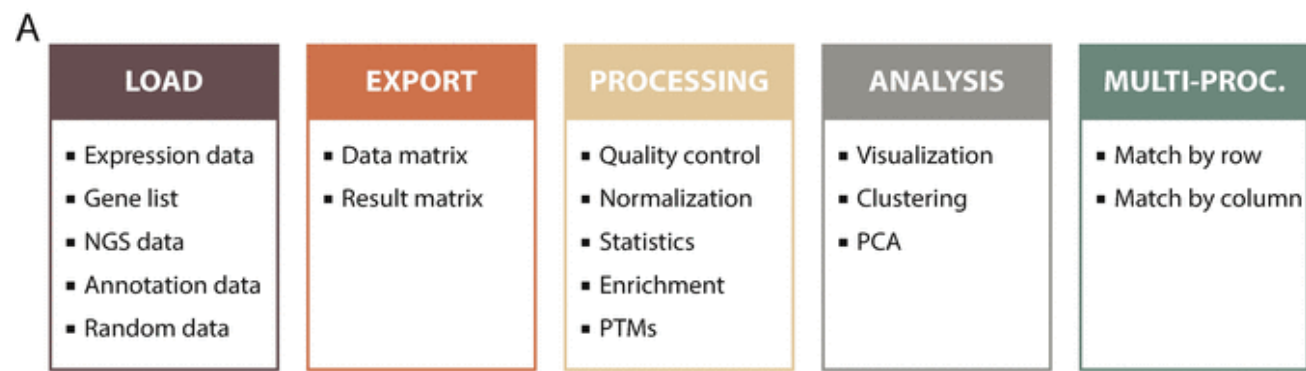
# Perseus



<https://maxquant.net/perseus/>

# Summary of interface

Five interfaces



**B**

|                              |    |    |    |    |    |    |
|------------------------------|----|----|----|----|----|----|
| Treatment                    | +  | +  | +  | -  | -  | -  |
| Technical replicates         | r1 | r2 | r3 | r1 | r2 | r3 |
| Numerical variable, e.g. BMI | 32 | 26 | 23 | 27 | 23 | 22 |

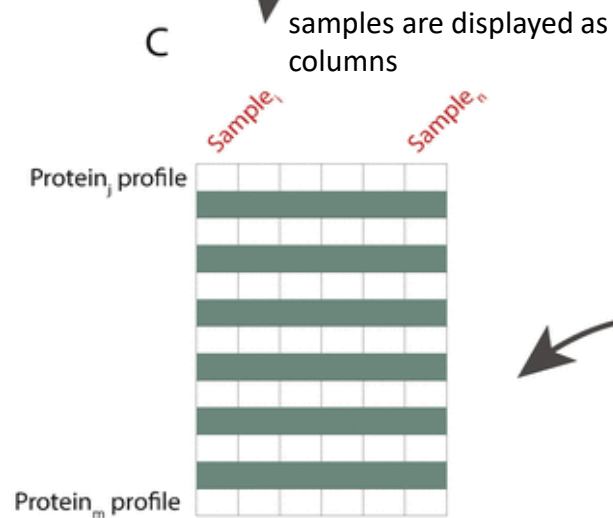
Annotation

Numerical

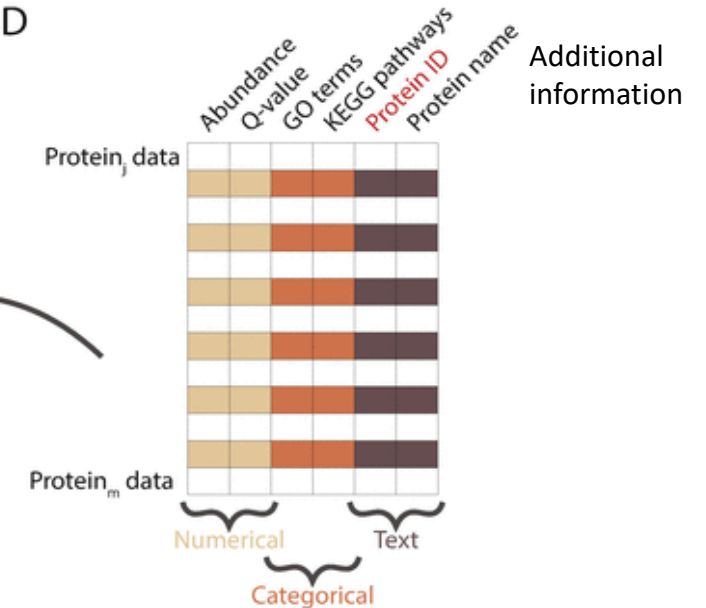
Experimental design is specified as annotation or numerical rows

Multiple annotation rows allow biological and technical replicates to be analyzed together

**C**



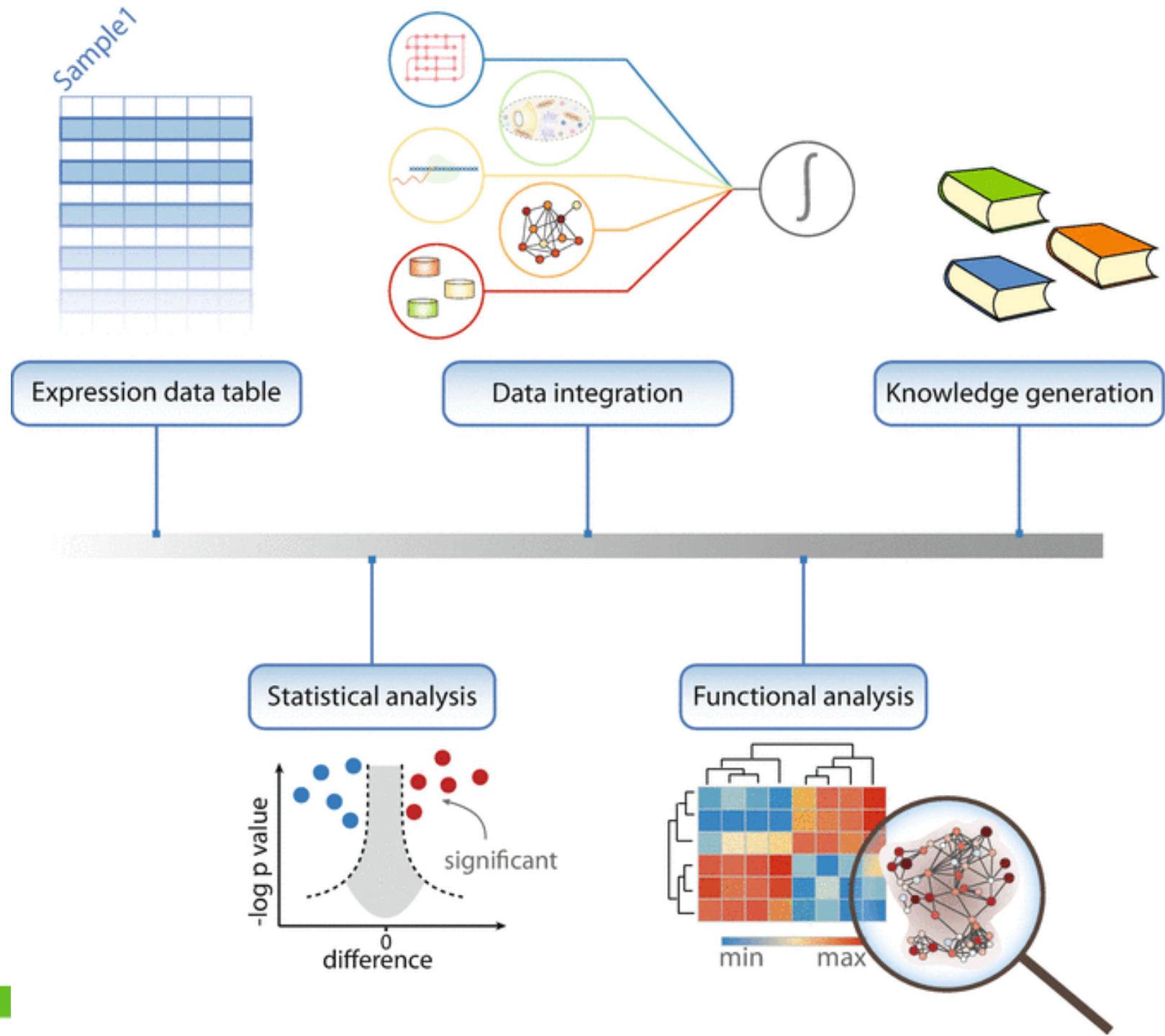
**D**



Augmented data matrix format



# A typical analysis workflow in Perseus





# Data Analysis using Perseus

1. Annotation
2. Filter / Extract
3. Expression profile / heatmap / cluster analysis
4. Functional enrichment analysis

This tutorial is based on Tyanova S., Cox J. (2018) Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. In: von Stechow L. (eds) Cancer Systems Biology. Methods in Molecular Biology, vol 1711. Humana Press, New York, NY  
[https://doi.org/10.1007/978-1-4939-7493-1\\_7](https://doi.org/10.1007/978-1-4939-7493-1_7)

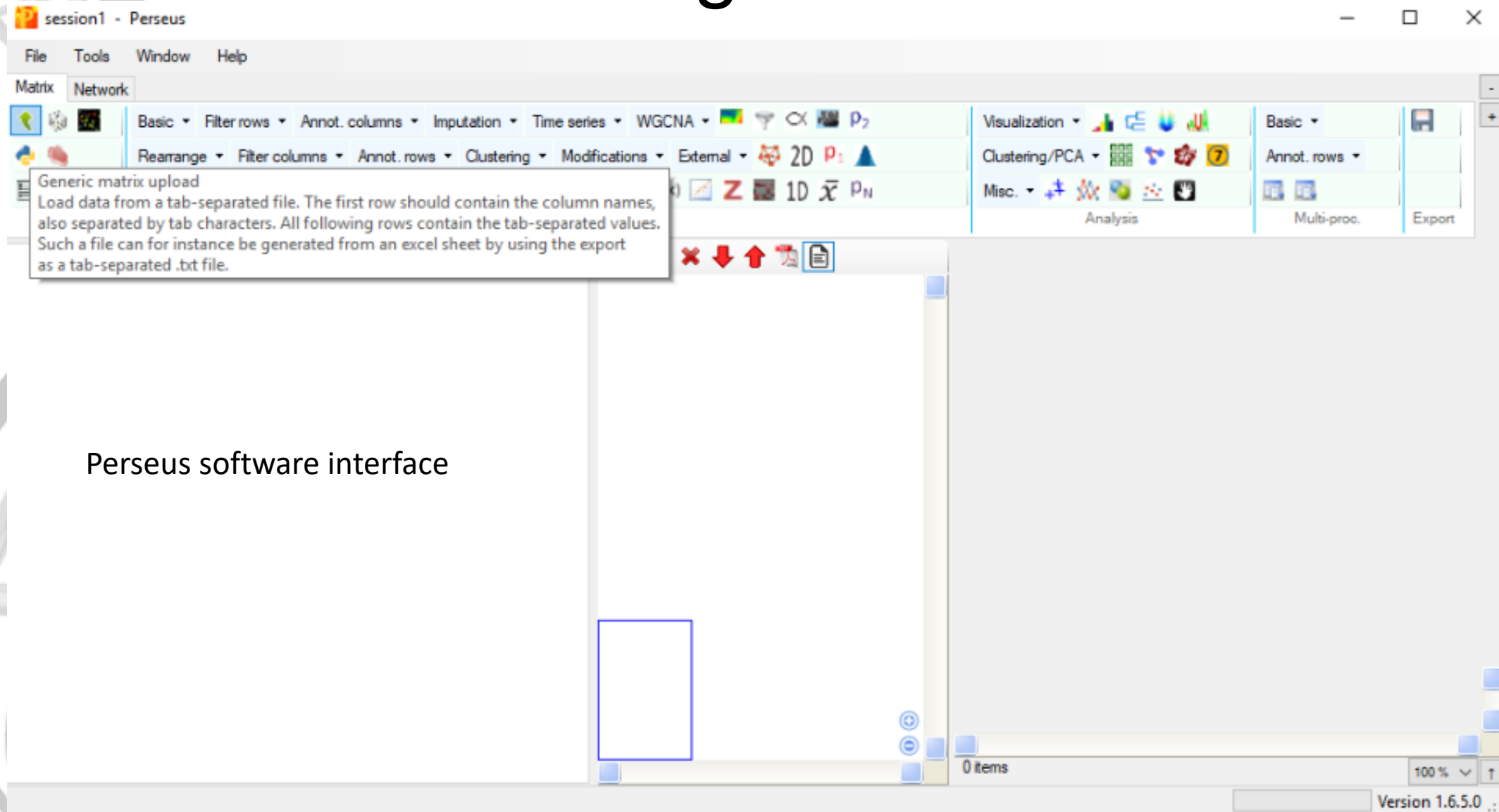


# Perseus Demo

- Loading the data
- Filtering
- Exploratory analysis
- Loading annotations
- Differential expression analysis
- Clustering & Profile plots
- Functional analysis



# Loading the data



Perseus software interface

### 3.1 Loading the Data

1. Go to the “Load” section in Perseus and click the “Generic matrix upload” button.
2. In the pop-up window, navigate to the file to be loaded (see **Note 2**).
3. Select all the expression columns and transfer them to the *Main* columns window (see **Note 3**). Select all additional numerical data that may be needed in the analysis and transfer them to the *Numerical* columns window. Make sure that the columns containing identifiers (e.g., protein IDs) are selected as *Text* columns. Click *ok*.

The screenshot displays the Perseus software interface. On the left, the 'InitialData' matrix is shown with columns A11H, A14H, A15H, and A16H. The 'Matrix' tab is active, showing a table with 8 rows and 5 columns. The 'Workflow panel' is in the center, showing a flow from 'Generic matrix u...' to 'InitialData'. The 'Matrix information' panel on the right provides details about the matrix, including the creator, origin, file, quality, and the number of rows, columns, and rows of various types.

| Type | Main      | Main      | Main      | Main    |
|------|-----------|-----------|-----------|---------|
| 1    | 0         | 0         | 0         | 0       |
| 2    | 0         | 380360... | 8777700   | 2682... |
| 3    | 6284300   | 5667100   | 332590... | 1494... |
| 4    | 240450... | 2260200   | 719530... | 2969... |
| 5    | 102760... | 344480... | 375810... | 9209... |
| 6    | 0         | 0         | 0         | 0       |
| 7    | 52412     | 0         | 2222100   | 0       |
| 8    | 194590... | 2852700   | 4892100   | 3488... |

Workflow panel

Matrix information

Get familiar with the Software and its five main sections: Load, Processing, Analysis, Multi-processing, and Export (see Fig. 2).

1. In the workflow panel, change the name of the data matrix from *matrix 1* to *InitialData* by right-clicking the node and changing the *Alternative name* box. Close the pop-up window. Explore the right-most panel of Perseus, which contains useful information such as number of main columns and number of rows.

The screenshot shows the 'Generic matrix upload' dialog box. The 'File' field contains the path 'C:\Users\Administrator\Goh\Prof\Society\MAPS\Workshop\2019\Materials\proteinGroups.txt'. The 'Main' section lists columns A11H, A14H, A15H, A16H, A19H, A30H, A31H, A34H, and A35H. The 'Numerical' section lists 'Q-value' and '# peptides'. The 'Categorical' section is empty. The 'Text' section lists 'Gene names', 'Protein names', and 'Protein IDs'. The 'Multi-numerical' section is empty. The 'Shorten main column names' checkbox is unchecked.

Generic matrix upload

File: C:\Users\Administrator\Goh\Prof\Society\MAPS\Workshop\2019\Materials\proteinGroups.txt

Main

- A11H
- A14H
- A15H
- A16H
- A19H
- A30H
- A31H
- A34H
- A35H

Numerical

- Q-value
- # peptides

Categorical

Text

- Gene names
- Protein names
- Protein IDs

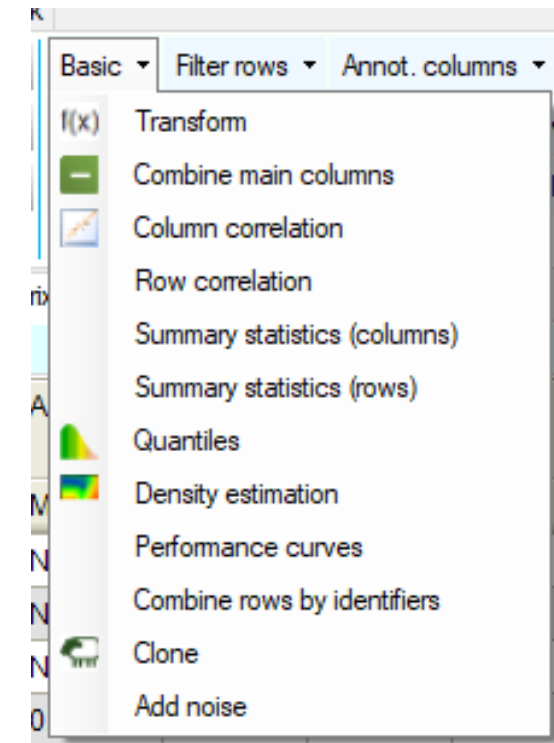
Multi-numerical

☐ Shorten main column names

# Data transformation

Transform the data to a logarithmic scale by going to “Processing → Basic → Transform” and specifying the transformation function (e.g.,  $\log_2(x)$ ).

In the “Processing” section, select the “Basic” menu and click on the “Summary statistics (columns)” button. Select all expression columns by transferring them to the right-hand side. Click *ok* and explore the new matrix.



## Summary statistics

InitialData matrix2 matrix3

Data

|      | Sum     | Mean    | Median  | Tukey biweight | Standa...<br>deviati... | Coeffic...<br>of variation | Median<br>absolute<br>deviati... | Full<br>width<br>half | Minimu... |
|------|---------|---------|---------|----------------|-------------------------|----------------------------|----------------------------------|-----------------------|-----------|
| Type | Main    | Main    | Main    | Main           | Main                    | Main                       | Main                             | Main                  | Main      |
| 1    | 96757.6 | 23.5592 | 23.5737 | 23.5626        | 2.6912                  | 0.1142...                  | 1.84678                          | 6.68063               | 14.5273   |
| 2    | 114150  | 23.0513 | 22.987  | 22.9965        | 2.65355                 | 0.1151...                  | 1.78852                          | 6.15427               | 14.4081   |
| 3    | 128974  | 23.7171 | 23.7599 | 23.7315        | 2.88448                 | 0.12162                    | 2.04989                          | 7.1232                | 13.011    |
| 4    | 116775  | 24.5481 | 24.5877 | 24.5492        | 2.83811                 | 0.1156...                  | 1.94825                          | 6.38805               | 14.9783   |
| 5    | 103881  | 24.4828 | 24.5774 | 24.5512        | 2.87305                 | 0.11735                    | 1.91152                          | 6.2185                | 15.1522   |
| 6    | 168764  | 23.9687 | 24.0155 | 24.0064        | 3.14415                 | 0.1311...                  | 2.23054                          | 8.25982               | 12.7439   |
| 7    | 134523  | 23.2056 | 23.2607 | 23.2178        | 3.11658                 | 0.1343...                  | 2.27908                          | 8.19807               | 13.7836   |

Generic matrix u...

InitialData

Transform

matrix2

Summary statisti...

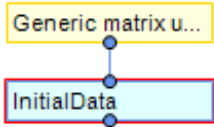
matrix3

matrix3

- InitialData
- Creator: Administrator
- 06/03/2019 17:29:01
- Origin: C:\Users\Administrato
- File: proteinGroups.txt
- Rows (88)
- Main columns (17)
- Categorical columns (0)
- String columns (1)
- Numerical columns (0)
- Multi-numerical columns (0)
- Categorical rows (0)
- String rows (0)
- Numerical rows (0)
- Multi-numerical rows (0)



# Filter

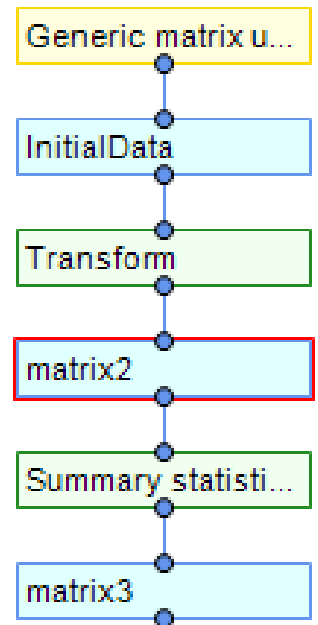


### 3.3 Filtering

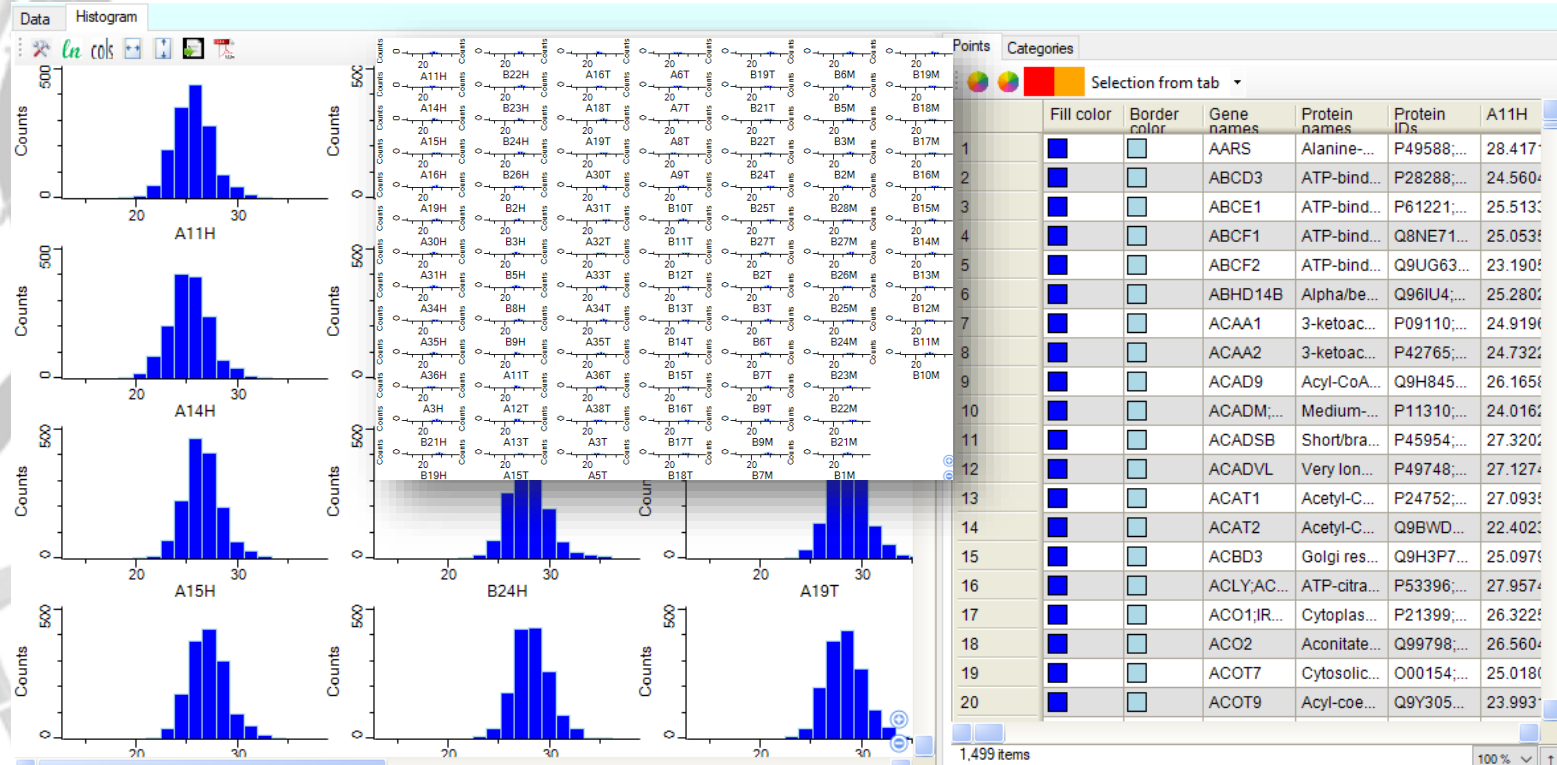
1. Use the workflow window to select the *InitialData* matrix data by clicking on it (see **Note 5**).
2. In the “Processing” section, go to the “Filter rows” menu and select “Filter rows based on valid values.” Change the *Min. valids* parameter to *Percentage* and keep the default value of 70% for the *Min. percentage of values* parameter. Click *ok*. Check how many protein groups were retained after the filtering (see **Note 6**).

- 4,083 items

1,499 items



# Exploratory analysis



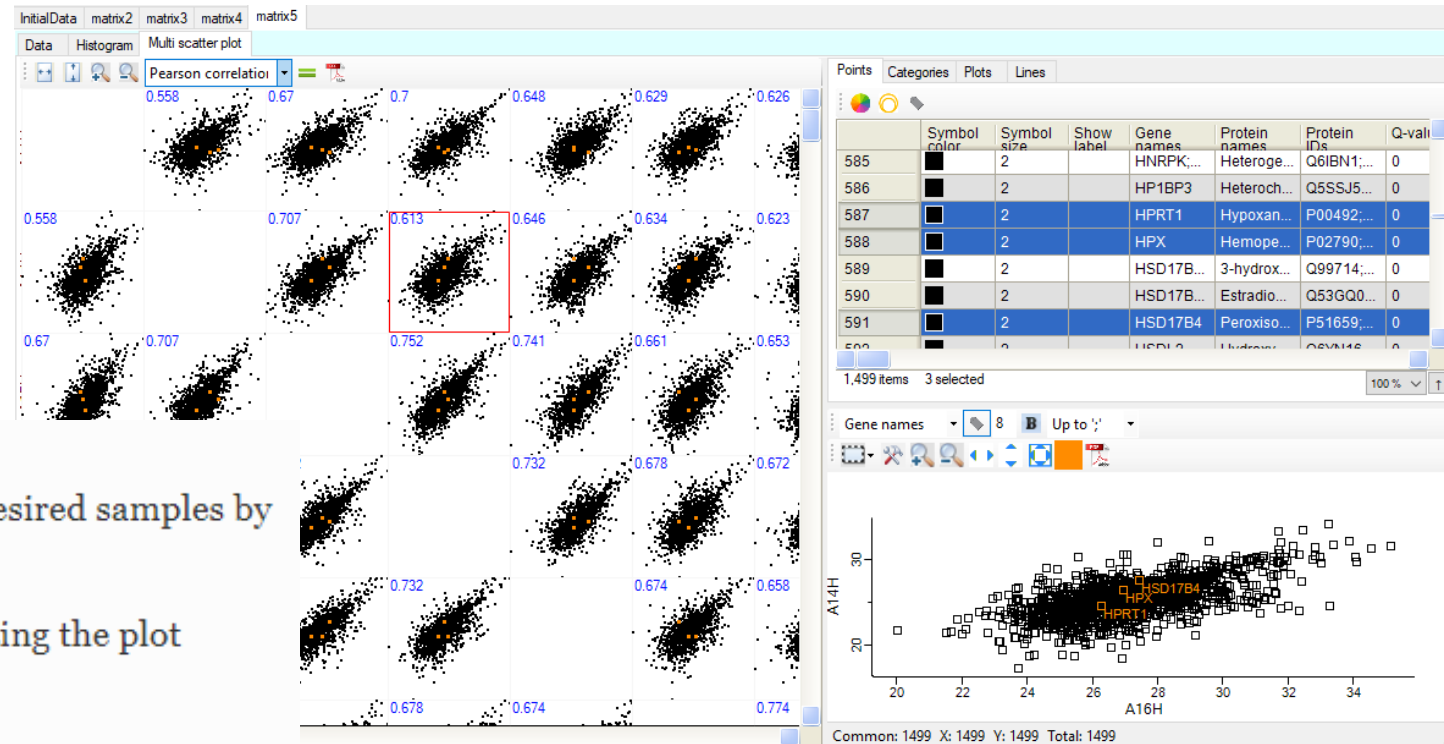
## Histogram

To visually inspect the data, go to “Analysis → Visualization → Histograms.” Select all the samples of interest by transferring them to the right-hand side. Click *ok*. Explore the visualization options in the Histogram panel by testing the functionality of each of the buttons (e.g., *Properties*, *Fit width*, *Fit height*). Click on the *pdf* button to export the plot (see [Note 1](#)).



# Exploratory analysis

## Multi-scatter plot



Switch the view to the “Data” tab.

Go to “Analysis → Visualization → Multi scatter plot.” Select the desired samples by transferring them to the right-hand side. Click *ok* (see Fig. 3).

Adjust the plot using the *Fit width* and *Fit height* options and resizing the plot window.

In the drop-down menu “Display in plots” in the plot window, select *Pearson correlation*.

Select a scatter plot by clicking on it. The selected plot will be shown in an enlarged view.

Select a number of proteins from the “Point” table on the right of the multi scatter plot and examine their position in all pairwise sample comparisons.

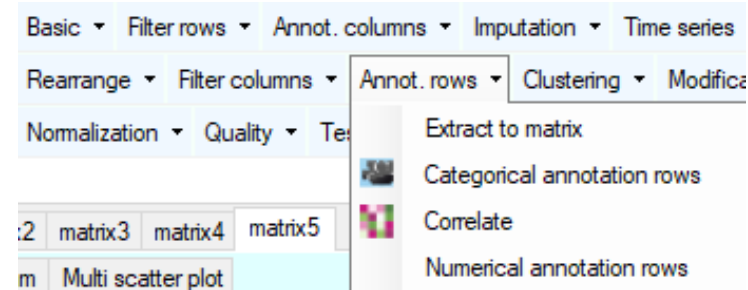
Switch back to the “Data” tab to continue with the analysis.



# Exploratory analysis

## Experimental Design

Go to “Processing → Annot. rows → Categorical annotation rows.” Use the *Create action* option to manually specify the experimental condition to which a sample belongs (i.e., indicate control versus stimulus, or different stages of a disease). All the samples belonging to one condition should have the same annotation. A new row will be added under the column names in the newly generated data matrix (see



## Categorical annotation

Abbreviation of clinical samples is as follows:

A- Lymph node negative case

B- Lymph node positive case

The prefix is followed by the serial number of the sample

H- Healthy duct

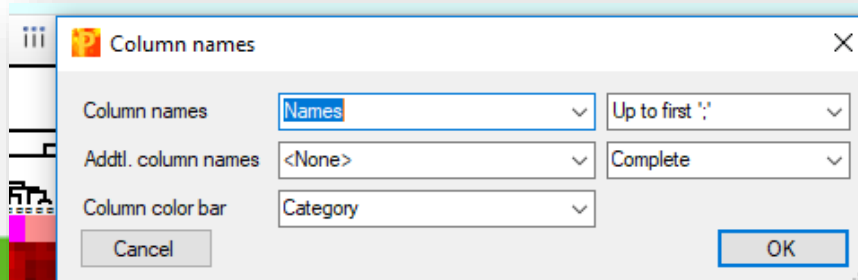
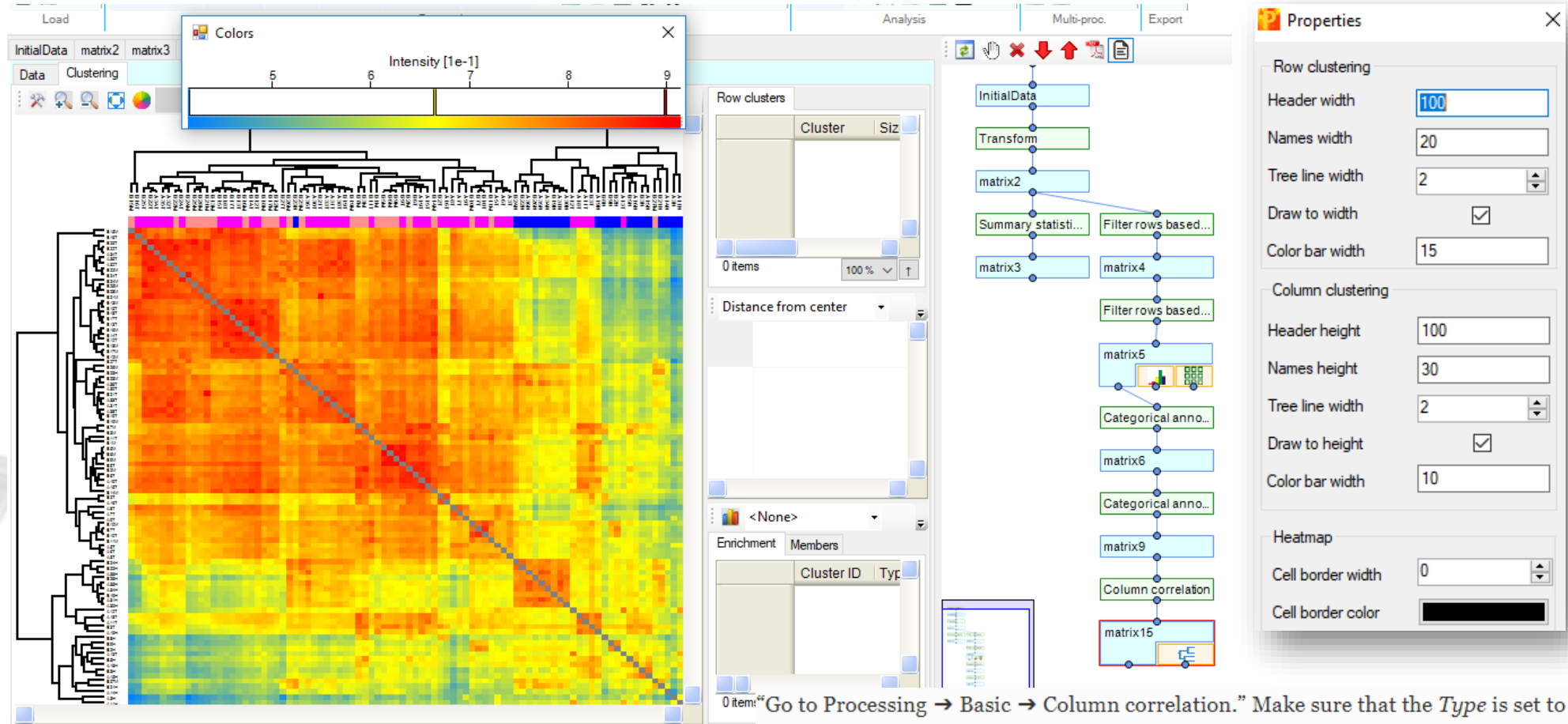
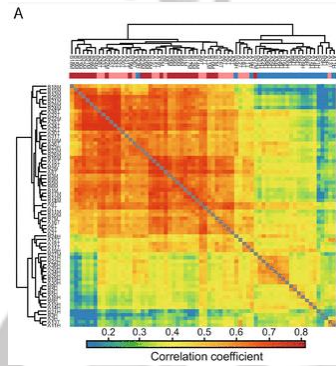
T- Primary tumor

M- Lymph node metastasis

|          | A11H      | A14H      | A15H      | A16H      | A19H      | A30H      | A31H      | A34H      | A35H      | A36H      | A3H       | B21H     | B19H     | B22H     |
|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|----------|----------|----------|
| Type     | Main      | Main      | Main      | Main      | Main      | Main      | Main      | Main      | Main      | Main      | Main      | Main     | Main     | Main     |
| Group1   | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Negati... | Positive | Positive | Positive |
| Category | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy   | Healthy  | Healthy  | Healthy  |
| 1        | 28.4171   | 27.7603   | 28.339    | 27.8714   | 29.0861   | 28.866    | 27.4809   | 27.0786   | 27.3449   | 27.3757   | 26.5142   | 28.4694  | 30.0924  | 29.0399  |
| 2        | 24.5604   | 23.9485   | 27.1361   | 27.5577   | 26.8517   | 27.1885   | 24.4723   | 24.0965   | 25.3176   | 25.8664   | 23.1134   | 25.7556  | 26.3849  | 26.6711  |
| 3        | 25.5133   | 25.5058   | 26.5467   | 27.1576   | 26.5711   | 27.7348   | 29.1806   | 26.3452   | 29.309    | 27.2445   | 25.4215   | 27.0525  | 28.2941  | 30.7804  |

# Exploratory analysis

## Hierarchical clustering



Go to Processing → Basic → Column correlation.” Make sure that the *Type* is set to *Pearson correlation*. The output table contains all pairwise correlations between the selected columns.

To visualize the sample correlations, go to “Analysis → Clustering/PCA → Hierarchical clustering.” Use the *Change color gradient* to set a continuous gradient

## Sample correlation

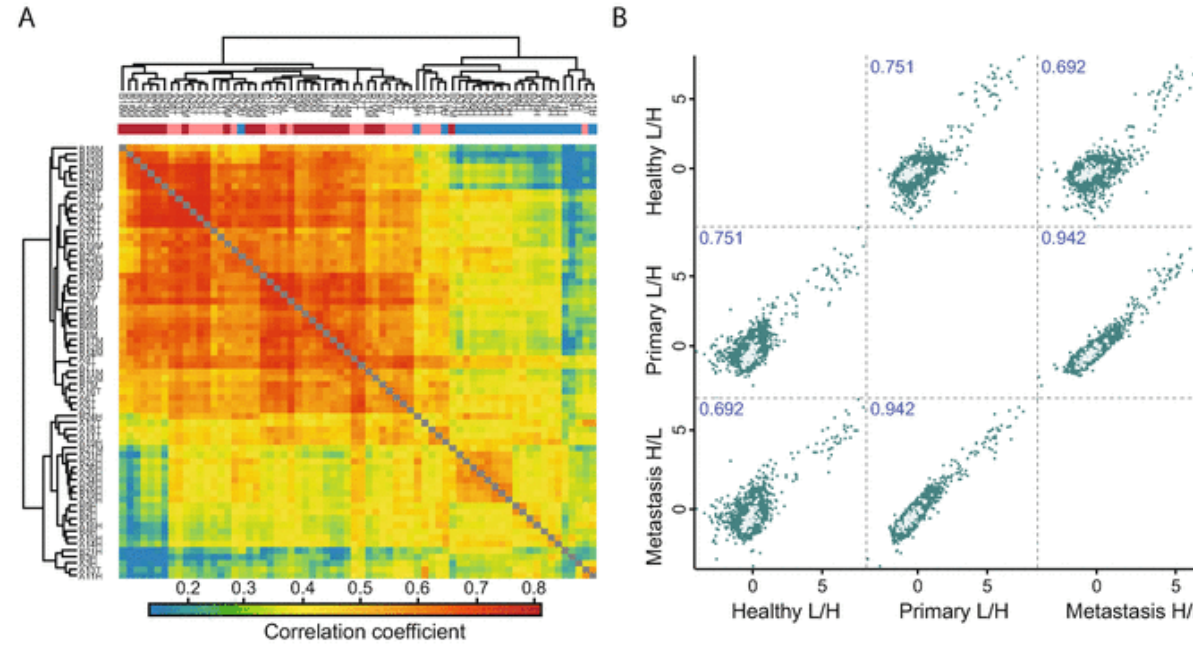
# Principal component analysis (PCA)



Explore the proteins driving this separation. In the loadings plot beneath the PCA, change the selection *Mode* to *rectangular selection*. Hold the left mouse key down and draw a rectangle around the dots in the upper right corner and then release the mouse. The selected proteins are highlighted in the table to the right and their labels are displayed in the plot.

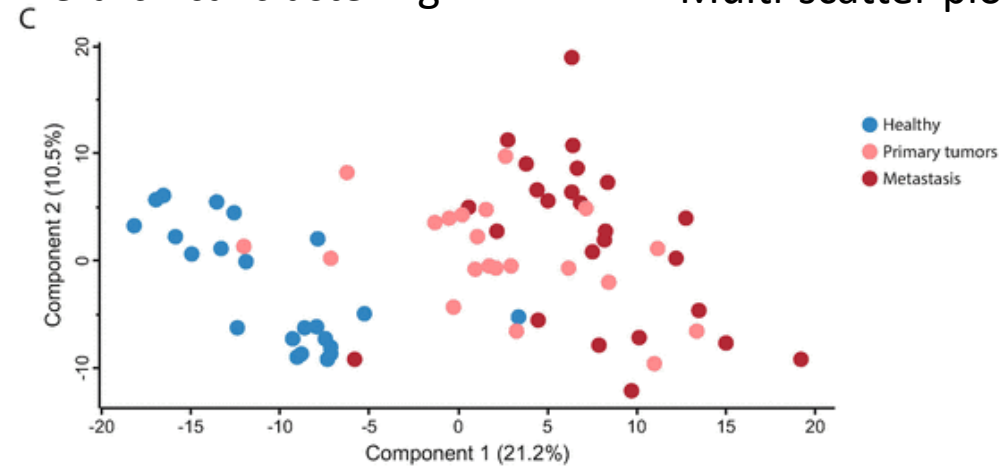


# Exploratory analysis



Hierarchical clustering

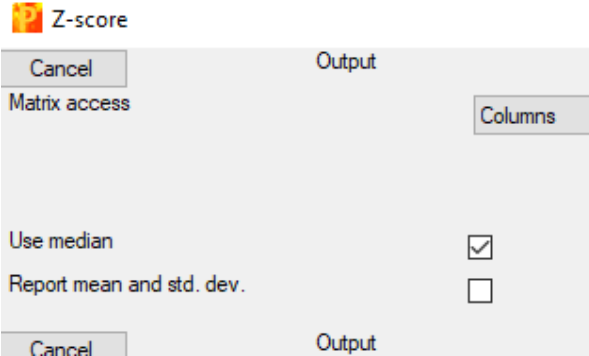
Multi-scatter plot



Principal component analysis (PCA)

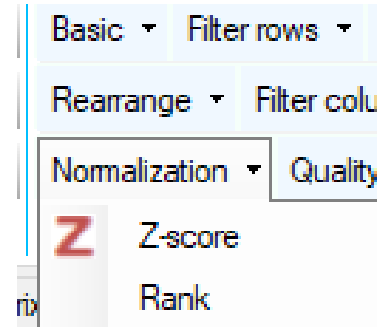


# Differential expression analysis



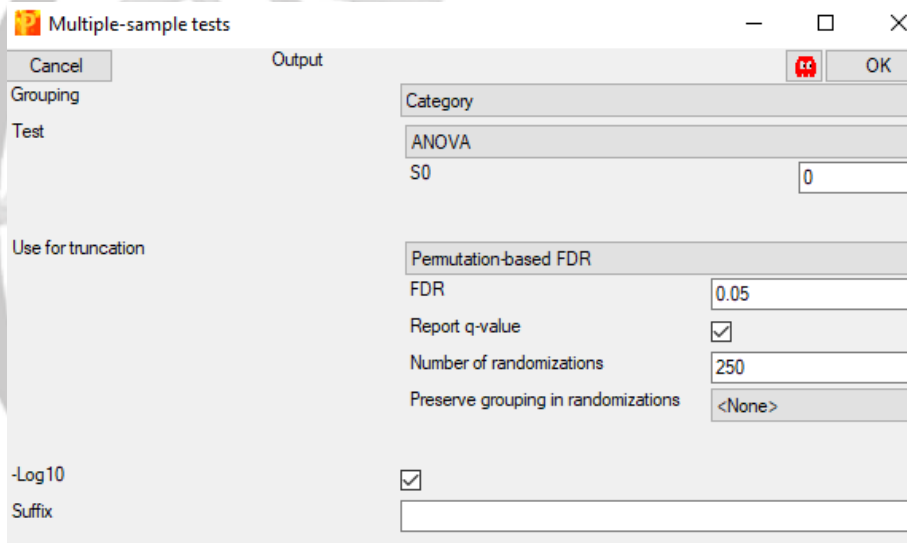
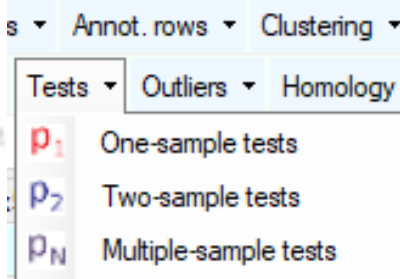
## Normalization

Go to “Processing → Normalization → Z-score.” Change the *Matrix access* parameter to *Columns* and select the *Use median* option.



## Differential Expression Analysis

Go to “Processing → Tests.” From the menu select the *Multiple-sample tests* as there are more than two conditions that are compared. The default parameters do not have to be changed.



Specify the categorical row that contains information about the experimental conditions of the samples that will be used in the differential analysis in the *Grouping* parameter.

Keep the default value of 0 for the *S0* parameter, to use the standard t-test statistic.

Change the parameter to use the modified test statistic approach described by Tusher et al. [15].

Select the multiple hypothesis testing correction method to be used by specifying the *Use for truncation* parameter (see **Note 12**, Fig. 4a).

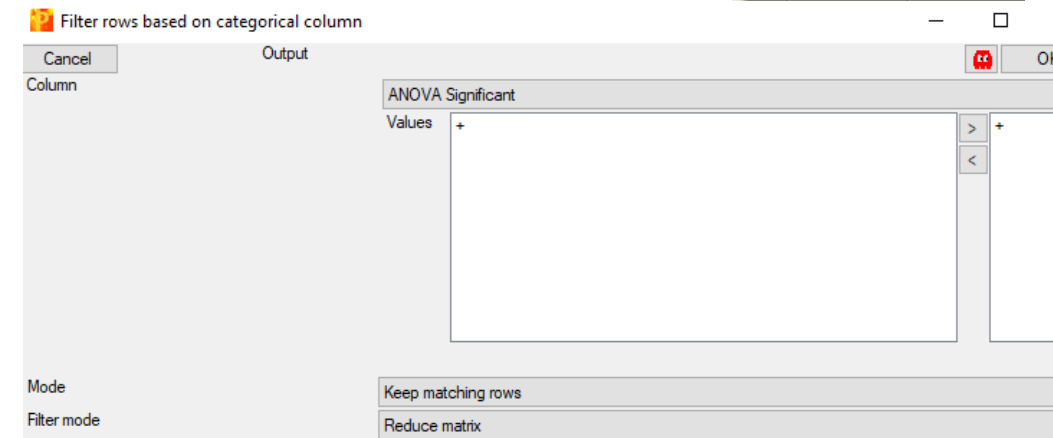
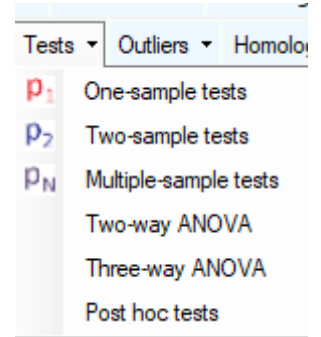
Specify if a suffix should be added to the output columns produced by Perseus. This option is relevant when multiple tests are conducted, e.g., with different parameter settings, as it helps to distinguish between them in the output table.

Inspect the output table. It contains three new columns: *ANOVA significant*, *-Log ANOVA p-value*, and *ANOVA q-value* (see **Note 13**).

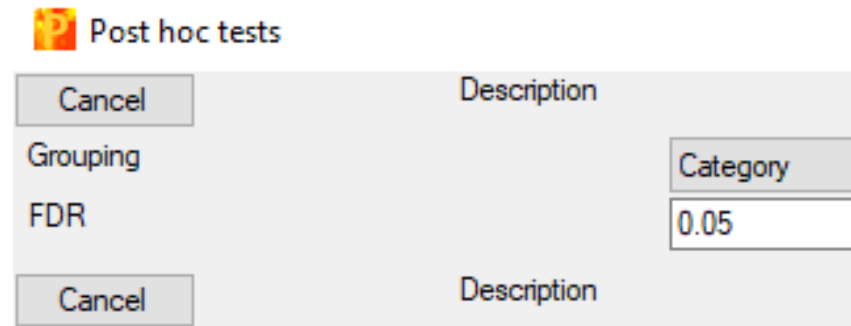
# Differential expression analysis (II)

Go to “Processing → Filter rows → Filter rows based on categorical column.” Set the *Column* parameter to *ANOVA Significant* and the *Mode* parameter to *Keep matching rows* to retain all differentially expressed proteins.

Go to “Processing → Tests → Post-hoc tests.” Set the *Grouping* parameter to the same grouping that was used for the ANOVA test (see Subheading 3.6, step 1) and the FDR to the desired threshold. Tukey’s honestly significant difference (THSD) is computed for all proteins and all pairwise comparisons and the significant hits within the corresponding pairs are marked (see Note 14, Fig. 4b).

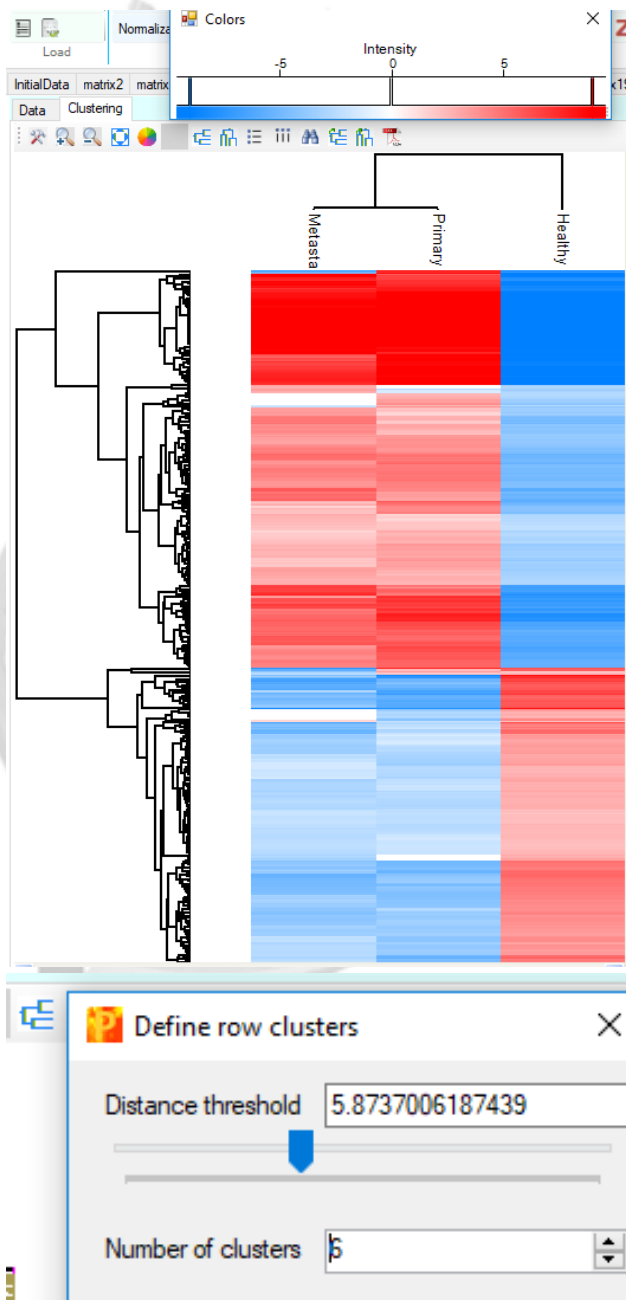


| Healthy  | Primary tumours | Metast... | C: ANOVA Significant | N: Q-value | N: # peptides | N: -Log ANOVA p value | N: ANOVA q-value | T: Gene names | T: Protein names | T: Protein IDs | T: Significant pairs      |
|----------|-----------------|-----------|----------------------|------------|---------------|-----------------------|------------------|---------------|------------------|----------------|---------------------------|
| Main     | Main            | Main      | Catego...            | Numeric    | Numeric       | Numeric               | Numeric          | Text          | Text             | Text           | Text                      |
| 2.76183  | -2.76183        | -2.57709  | +                    | 0          | 12            | 3.07782               | 0.0200...        | AARS          | Alanine...       | P4958...       | Healthy_Metastasis;Hea... |
| 3.94078  | -3.60059        | -3.94078  | +                    | 0          | 7             | 4.35561               | 0.0035...        | ABCE1         | ATP-b...         | P6122...       | Healthy_Metastasis;Hea... |
| 3.40231  | -1.65417        | -3.40231  | +                    | 0          | 4             | 3.21502               | 0.0172...        | ABCF1         | ATP-b...         | Q8NE7...       | Healthy_Metastasis;Hea... |
| 7.40451  | -7.40451        | -3.24112  | +                    | 0          | 4             | 7.06613               | 3.1746...        | ABCF2         | ATP-b...         | Q9UG...        | Healthy_Primary tumour... |
| -4.61701 | 4.61701         | 4.24983   | +                    | 0          | 22            | 5.12048               | 0.0011...        | ABHD1...      | Alpha/...        | Q96IU...       | Primary tumours_Health... |
| 3.43283  | -2.48767        | -3.43283  | +                    | 0          | 14            | 3.4876                | 0.0123...        | ACSF2         | Acyl-C...        | Q96CM...       | Healthy_Metastasis;Hea... |





# Clustering



Go to “Analysis → Clustering/PCA → Hierarchical clustering.” Keep the default parameters and click *ok*.

Inspect the resulting heatmap and the relationship between the groups and the proteins.

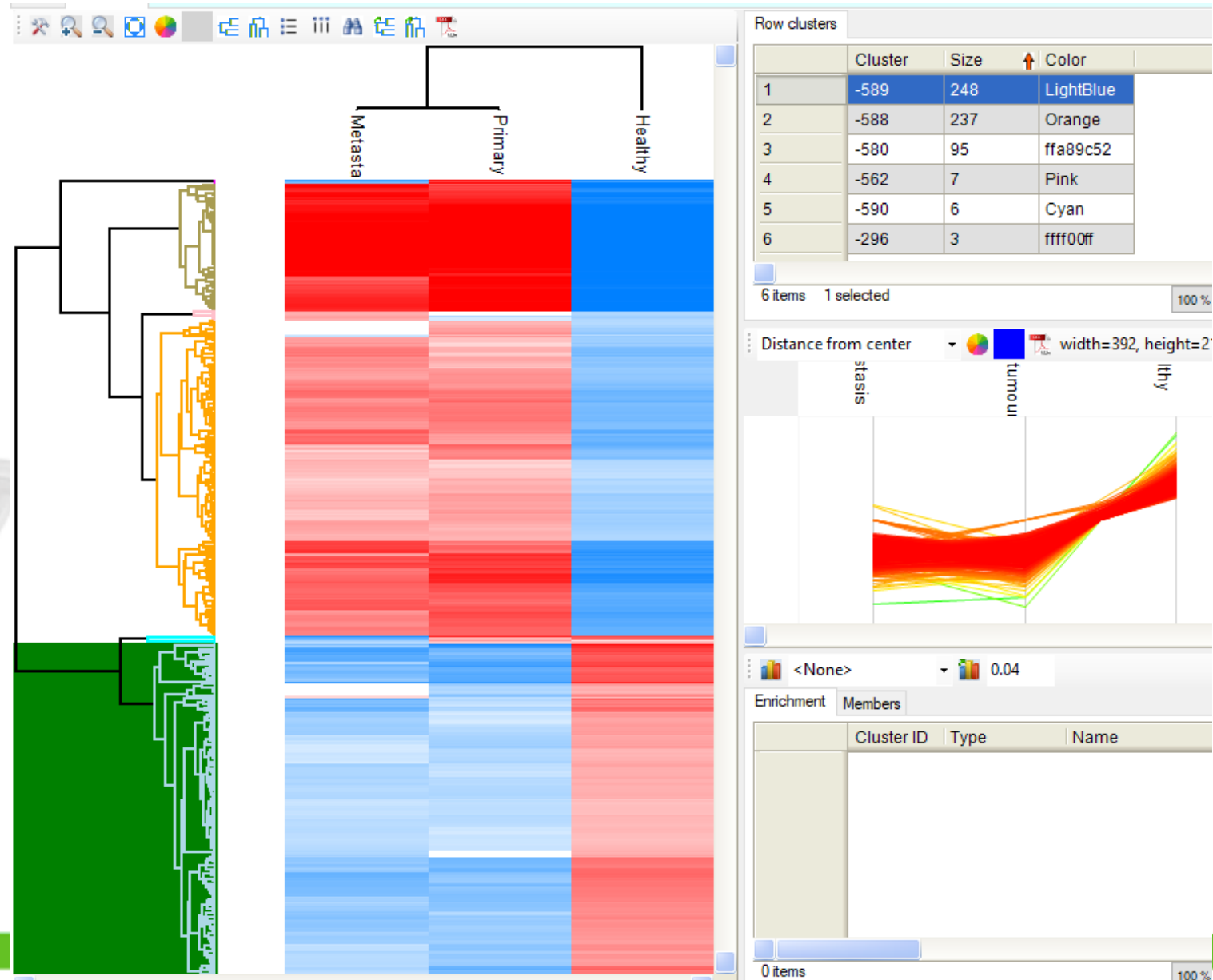
Click on the *Change color gradient* button in the button ribbon above the heatmap to examine the color scale usage (red means high and green low expression) and to modify them.

Click on several node junctions in the protein tree that represent potentially interesting clusters of proteins (i.e., upregulation in a certain experimental condition). The selected clusters are highlighted and appear in the “Row clusters” table displayed to the right of the heatmap (see **Note 15**).

Inspect the different profile plots as you navigate through the different clusters in the table. Change the color by modifying the *Color scale* and export the profile plots by clicking on the *Export image* button (see Fig. 5).

From the ribbon menu in the heat map view, click on the *Export row clustering* button to add the cluster information to a new data matrix.

# Profile plots



# Functional analysis

Go to “Multi-proc. → Matching rows by name.” Both *Base* and *Other* matrices point to the last matrix.

Click on *Base matrix* and then in the workflow window select the data matrix that was generated before filtering for ANOVA significant.

In the pop-up window set *Matching column in matrix 1* and *2* to a common identifier (e.g., *Protein IDs*).

In the categorical columns section, transfer the category *Cluster* to the right hand-side. Click *ok*

**Matching rows by name**  
The base matrix is copied. Rows of the second matrix are associated with rows of the base matrix via matching expressions in a textual column from each matrix. Selected columns of the second matrix are attached to the first matrix. If exactly one row of the second matrix corresponds to a row of the base matrix, values are just copied. If more than one row of the second matrix matches to a row of the first matrix, the corresponding values are averaged (actually the median is taken) for numerical and expression columns and concatenated for textual and categorical columns.

**Matching rows by name**

Cancel Description

Matching column in table 1 Gene names

Matching column in table 2 Gene names

Use additional column pair ☐

Join style

Ignore case ☐

Add indicator ☐

Add original row numbers ☐

Copy main columns

Combine copied main values

Copy categorical columns

Left

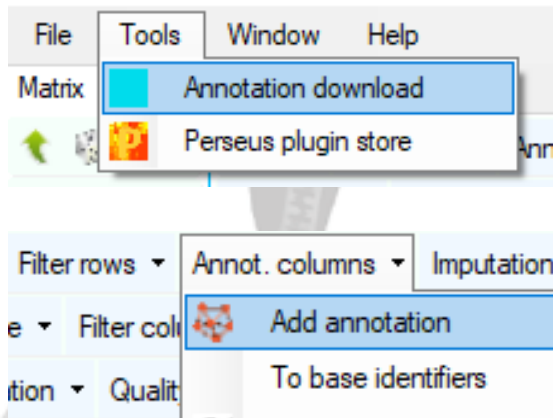
Metastasis  
Primary tumours  
Healthy

Median

ANOVA Significant  
Cluster

Cluster

# Loading annotations



Go to the drop-down menu indicated with a white arrow at the top left corner of Perseus and select “Annotation download.”

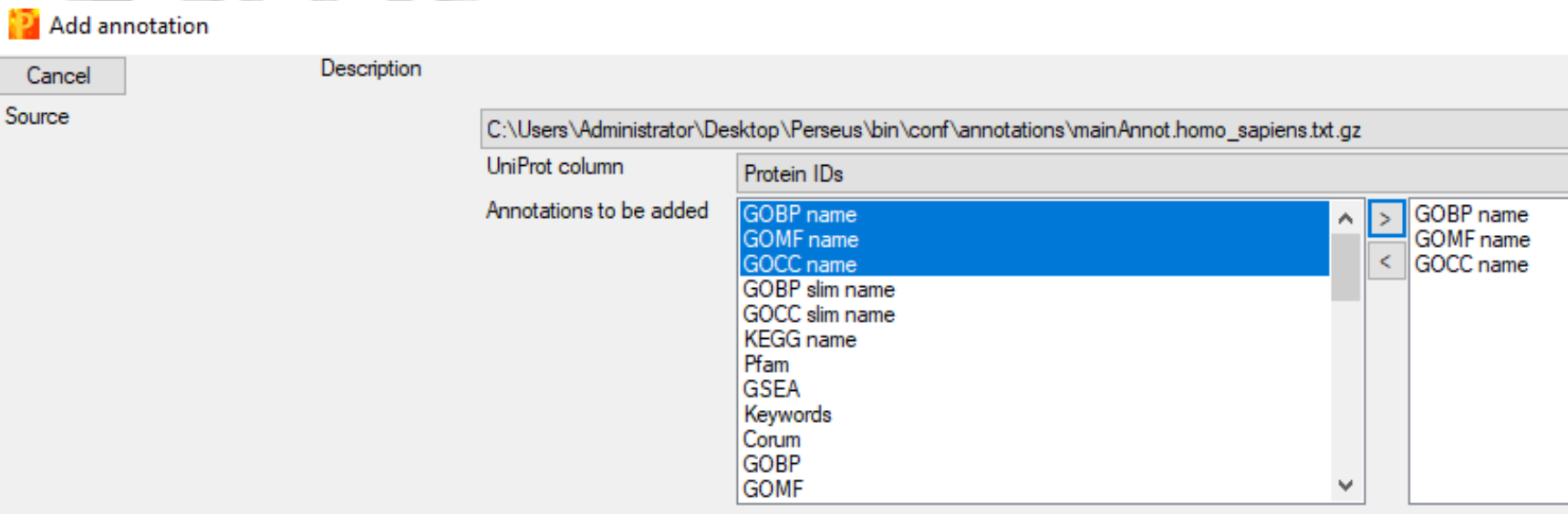
Click on the link in the pop-up window. Select the appropriate annotation file (e.g., “PerseusAnnotation → FrequentlyUsed → mainAnnot.homo\_sapiens.txt.gz,” if the organism of interest is *homo sapiens*).

Download the file to the *Perseus/conf/annotations* folder.

Go to “Processing → Annot. columns → Add annotation.” Select the file from the previous step as a *Source*.

Set the *UniProt column* parameter to the column that contains UniProt identifiers. These identifiers will be used for overlaying the annotation data with the expression matrix (e.g., *Protein IDs*).


Select several categories of interest to be overlaid with the main matrix and move them to the right-hand side. Click *ok*.



# Functional analysis

## Fisher's exact test

Go to "Processing → Annot. columns → Fisher exact test." Change the *Column* parameter to *Cluster* and click *ok*. The resulting table contains information about all annotation categories that were found to be significantly enriched or depleted using a and multiple hypotheses correction

 Fisher exact test

Cancel Output

Input type: Categorical column

Column: Cluster

Use for truncation: Benjamini-Hochberg FDR

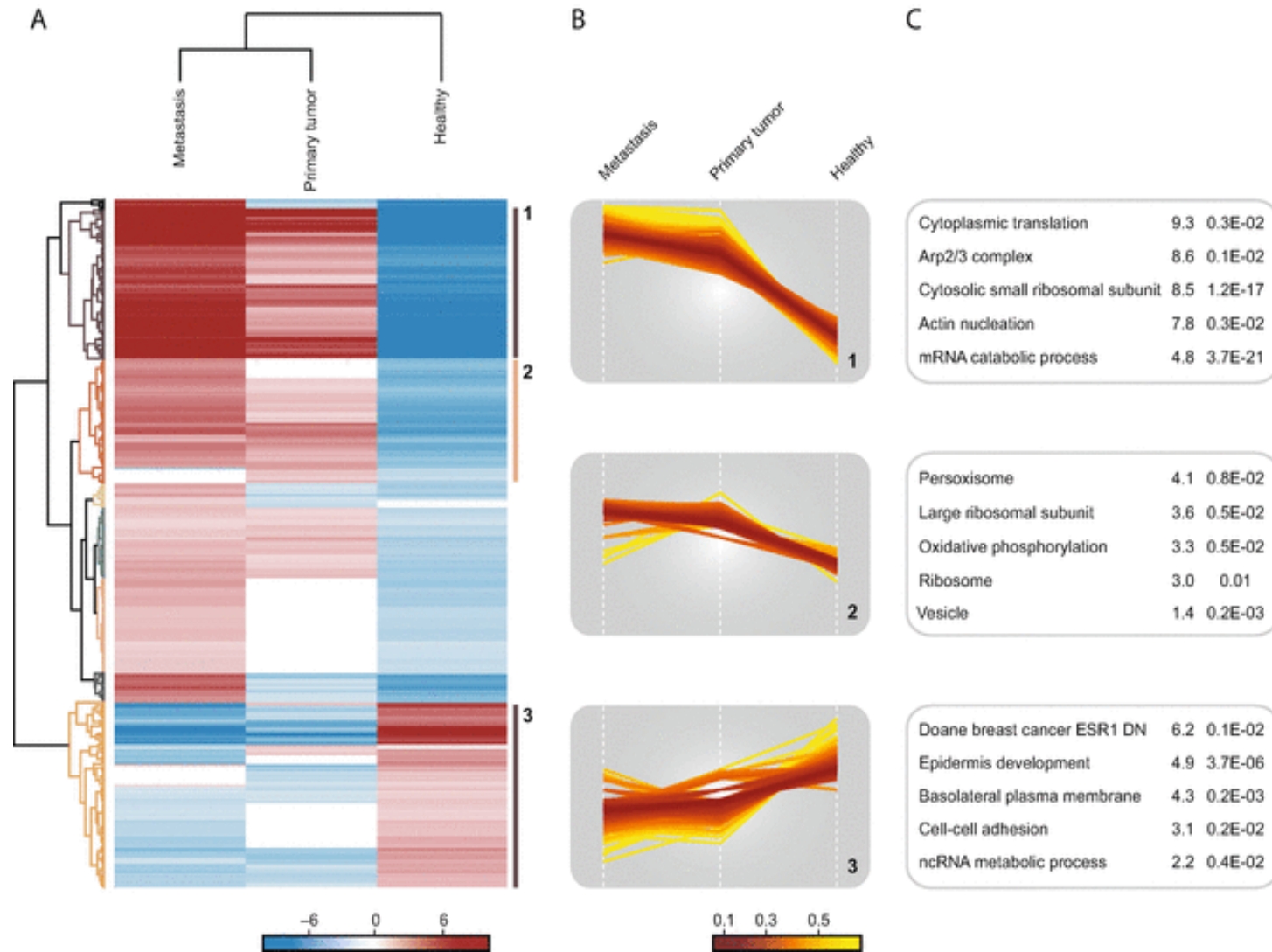
Threshold value: 0.05

Relative enrichment: <None>

| Data |                       |                       |                        |                                      |                  |                       |                      |                       |                        |               |                          |  |
|------|-----------------------|-----------------------|------------------------|--------------------------------------|------------------|-----------------------|----------------------|-----------------------|------------------------|---------------|--------------------------|--|
|      | C: Selec...<br>column | C: Selection<br>value | C: Catego...<br>column | C: Category value                    | N: Total<br>size | N: Selecti...<br>size | N: Catego...<br>size | N: Interse...<br>size | N: Enrich...<br>factor | N: P<br>value | N: Benj.<br>Hoch.<br>FDR |  |
| Type | Categ...              | Category              | Catego...              | Category                             | Numeric          | Numeric               | Numeric              | Numeric               | Numeric                | Numeric       | Numeric                  |  |
| 1    | Cluster               | Cluster -590          | GOCC...                | Golgi lumen                          | 1499             | 6                     | 2                    | 2                     | 249.83                 | 1.336E...     | 0.0020...                |  |
| 2    | Cluster               | Cluster -590          | GOBP...                | glycosaminoglycan biosynthet...      | 1499             | 6                     | 3                    | 2                     | 166.56                 | 3.9973...     | 0.01383                  |  |
| 3    | Cluster               | Cluster -590          | GOCC...                | fibrillar collagen                   | 1499             | 6                     | 3                    | 2                     | 166.56                 | 3.9973...     | 0.0047...                |  |
| 4    | Cluster               | Cluster -590          | GOBP...                | aminoglycan biosynthetic proc...     | 1499             | 6                     | 3                    | 2                     | 166.56                 | 3.9973...     | 0.01383                  |  |
| 5    | Cluster               | Cluster -590          | GOMF...                | extracellular matrix structural c... | 1499             | 6                     | 3                    | 2                     | 166.56                 | 3.9973...     | 0.0122...                |  |
| 6    | Cluster               | Cluster -296          | GOCC...                | proteinaceous extracellular ma...    | 1499             | 3                     | 8                    | 2                     | 124.92                 | 7.4516...     | 0.0076...                |  |
| 7    | Cluster               | Cluster -590          | GOBP...                | glycosaminoglycan catabolic p...     | 1499             | 6                     | 5                    | 2                     | 99.933                 | 0.0001...     | 0.0384...                |  |
| 8    | Cluster               | Cluster -296          | GOMF...                | heparin binding                      | 1499             | 3                     | 10                   | 2                     | 99.933                 | 0.0001...     | 0.0315...                |  |
| 9    | Cluster               | Cluster -590          | GOBP...                | aminoglycan catabolic process        | 1499             | 6                     | 5                    | 2                     | 99.933                 | 0.0001...     | 0.0384...                |  |



# Functional analysis



Hierarchical clustering

Profile plots

Functional enrichment analysis



# References & Learning Resources

- Tyanova S., Cox J. (2018) Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. In: von Stechow L. (eds) Cancer Systems Biology. Methods in Molecular Biology, vol 1711. Humana Press, New York, NY  
[https://doi.org/10.1007/978-1-4939-7493-1\\_7](https://doi.org/10.1007/978-1-4939-7493-1_7)
  - Pozniak Y, Balint-Lahat N, Rudolph JD, Lindskog C, Katzir R, Avivi C, Ponten F, Ruppén E, Barshack I, Geiger T (2016) System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. Cell Syst 2(3):172–184.  
<https://doi.org/10.1016/j.cels.2016.02.001>
  - [The Perseus computational platform for comprehensive analysis of \(prote\)omics data](#) *Nat. Methods* 2016.
- 
- ❑ <http://coxdocs.org/doku.php?id=perseus:user:tutorials>
  - ❑ [MaxQuant youtube channel](#)

# Happy Trying!

