

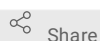
Sep 02, 2021

# Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research

Book Chapter

Juergen Cox, Stefka Tyanova

1 Works for me



Share

[dx.doi.org/10.17504/protocols.io.bnk7mczn](https://dx.doi.org/10.17504/protocols.io.bnk7mczn)

Springer Nature Books

satyavati Kharde

## ABSTRACT

Mass spectrometry-based proteomics is a continuously growing field marked by technological and methodological improvements. Cancer proteomics is aimed at pursuing goals such as accurate diagnosis, patient stratification, and biomarker discovery, relying on the richness of information of quantitative proteome profiles. Translating these high-dimensional data into biological findings of clinical importance necessitates the use of robust and powerful computational tools and methods. In this chapter, we provide a detailed description of standard analysis steps for a clinical proteomics dataset performed in Perseus, a software for functional analysis of large-scale quantitative omics data.

## ATTACHMENTS

Tyanova-  
Cox2018\_Protocol\_Perseu  
sABioinformaticsPlatform.  
pdf

## DOI

[dx.doi.org/10.17504/protocols.io.bnk7mczn](https://dx.doi.org/10.17504/protocols.io.bnk7mczn)

## EXTERNAL LINK

[https://link.springer.com/protocol/10.1007/978-1-4939-7493-1\\_7](https://link.springer.com/protocol/10.1007/978-1-4939-7493-1_7)

## PROTOCOL CITATION

Juergen Cox, Stefka Tyanova 2021. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.bnk7mczn>

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

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)

## KEYWORDS

Perseus, Software, Omics data analysis, Translational bioinformatics, Cancer proteomics

## LICENSE

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited






## CREATED

Oct 20, 2020

LAST MODIFIED

Sep 02, 2021

OWNERSHIP HISTORY

- Oct 20, 2020  Megan Freund
- Feb 03, 2021  Anita Broellochs protocols.io
- Jul 05, 2021  Emma Ganley protocols.io
- Aug 24, 2021  Satyavati Kharde
- Aug 26, 2021  satyavati Kharde

PROTOCOL INTEGER ID

43391

GUIDELINES

References

1. Mann M, Kulak NA, Nagaraj N, Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* 49(4):583–590. <https://doi.org/10.1016/j.molcel.2013.01.029>
2. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* 7(5):383–385. <https://doi.org/10.1038/nmeth.1446>
3. Shenoy A, Geiger T (2015) Super-SILAC: current trends and future perspectives. *Expert Rev Proteomics* 12(1):13–19. <https://doi.org/10.1586/14789450.2015.982538>
4. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13(9):2513–2526. <https://doi.org/10.1074/mcp.M113.031591>
5. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, Rodland KK, Townsend RR, Kinsinger C, Mesri M, Rodriguez H, Liebler DC, Clinical Proteomic Tumor Analysis C (2013) Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* 3(10):1108–1112. <https://doi.org/10.1158/2159-8290.CD-13-0219>
6. Hanash S, Taguchi A (2010) The grand challenge to decipher the cancer proteome. *Nat Rev Cancer* 10(9):652–660. <https://doi.org/10.1038/nrc2918>
7. Wisniewski JR, Dus-Szachniewicz K, Ostasiewicz P, Ziolkowski P, Rakus D, Mann M (2015) Absolute proteome analysis of colorectal mucosa, adenoma, and cancer reveals drastic changes in fatty acid metabolism and plasma membrane transporters. *J Proteome Res* 14(9):4005–4018. <https://doi.org/10.1021/acs.jproteome.5b00523>
8. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC, Nci C (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* 513(7518):382–387. <https://doi.org/10.1038/nature13438>
9. Iglesias-Gato D, Wikstrom P, Tyanova S, Lavalley C, Thysell E, Carlsson J, Hagglof C, Cox J, Andren O, Stattin P, Egevad L, Widmark A, Bjartell A, Collins CC, Bergh A, Geiger T, Mann M, Flores-Morales A (2016) The proteome of primary prostate cancer. *Eur Urol* 69(5):942–952. <https://doi.org/10.1016/j.eururo.2015.10.053>
10. Deeb SJ, Tyanova S, Hummel M, Schmidt-Supprian M, Cox J, Mann M (2015) Machine learning based classification of diffuse large B-cell lymphoma patients by their protein expression profiles. *Mol Cell Proteomics* 14(11):2947–2960. <https://doi.org/10.1074/mcp.M115.050245>
11. Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T (2016) Proteomic maps of breast cancer subtypes. *Nat Commun* 7:10259. <https://doi.org/10.1038/ncomms10259>
12. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, Kawaler E, Mundt F, Krug K, Tu Z, Lei JT, Gatta ML, Wilkerson M, Perou CM, Yellapantula V, Huang KL, Lin C, McLellan MD, Yan P, Davies SR, Townsend RR, Skates SJ, Wang J, Zhang B, Kinsinger CR, Mesri M, Rodriguez H, Ding L, Paulovich AG, Fenyo D, Ellis MJ, Carr SA, Nci C (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534(7605):55–62. <https://doi.org/10.1038/nature18003>
13. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6):520–525

14. Lazar C, Gatto L, Ferro M, Bruley C, Burger T (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J Proteome Res* 15(4):1116–1125. <https://doi.org/10.1021/acs.jproteome.5b00981>
15. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9):5116–5121. <https://doi.org/10.1073/pnas.091062498>
16. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* 57:289–300
17. Fisher RA (1922) On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *J R Stat Soc* 85:87–94. <https://doi.org/10.2307/2340521>
18. Cox J, Mann M (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* 13(Suppl 16):S12. <https://doi.org/10.1186/1471-2105-13-S16-S12>
19. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13(9):731–740. <https://doi.org/10.1038/nmeth.3901>
20. 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>
21. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367–1372. <https://doi.org/10.1038/nbt.1511>
22. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* 473(7347):337–342. <https://doi.org/10.1038/nature10098>

#### MATERIALS TEXT

##### Software Download and Installation

Written in C#, Perseus achieves optimal performance when run on Windows operating systems. The latest versions require 64 bit system and .NET Framework 4.5 to be installed on the same computer. To use the software on MacOS set up BootCamp and optionally in addition Parallels. Registration and acceptance of the Software License Agreement are required prior to downloading Perseus from the official website: <http://www.coxdocs.org/doku.php?id=perseus:start>. Once the download has finished, decompress the folder, locate the Perseus.exe file, and double-click it to start the program.

##### Data Files

In the subsequent analysis, we used a subset of the data measured by Pozniak et al. [20]. The authors provide a genome-wide proteomic analysis of progression of breast cancer in patients by studying major differences at the proteome level between healthy, primary tumor, and metastatic tissues. The data were measured as ratios between an optimized heavy-labeled mix of cell lines representing different breast cancer stages and the patient proteome [2]. This constitutes an accurate relative quantification approach used especially in the analysis of clinical samples. Peptide and protein identification and quantification was performed using the MaxQuant suite for the analysis of raw mass spectrometry data [21] at peptide spectrum match and protein false discovery rate of 1%. The subset used in this protocol contains proteome profiles of 22 healthy, 21 lymph node negative, and 25 lymph node metastatic tissue samples and spans over 10,000 protein groups and can be found in the proteinGroups.txt file provided as supplementary material to the Pozniak et al. study.

Proteins with shared peptides that cannot be distinguished based on the peptides identified in a bottom-up proteomics approach are often reported together as a protein group [21].

#### SAFETY WARNINGS

For hazard information and safety warnings, please refer to the SDS (Safety Data Sheet).

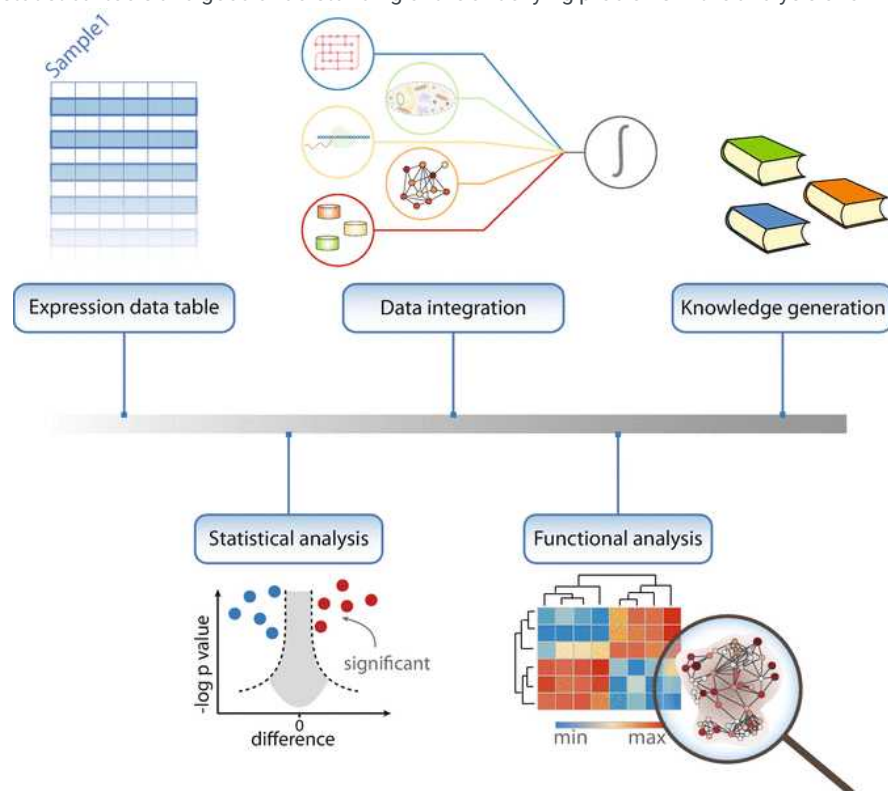
#### BEFORE STARTING

##### Introduction

High-resolution mass spectrometry-based proteomics, aided by computational sciences, is continuously pushing the boundaries of systems biology. Obtaining highly accurate quantitative proteomes on a genome-wide scale is

becoming feasible within realistic measurement times [1]. Similar to the clinical goals of genomics and transcriptomics to provide a deeper understanding of a certain disease that goes beyond the standard clinical parameters of cancer diagnosis, proteomics offers a comprehensive view of the molecular players in a cell at a particular moment and in a specific state [1]. The maturation of the technology together with the development of suitable methods for quantification of human tissue proteomes [2,3,4] has opened new doors for employing proteomics in medical applications and is shaping the growing field of clinical proteomics [5,6]. Following these advances, proteomic approaches have been used to address multiple clinical questions in the context of various cancer types. The major area of application is the profiling of cancer-relevant tissues—including the proteomes of colorectal cancer [7,8] and prostate cancer [9], as well as the subtyping of lymphoma [10] and breast cancer [11,12] patients. Although proteomics has become an extremely powerful approach for studying biomedical questions, offering unique advantages compared to other omics techniques, the functional interpretation of the vast amounts of data of a typical proteomics experiment often poses analytical challenges to the biological domain experts.

The aim of data analysis is to translate large amounts of proteomic data that cover numerous samples, conditions and time points into structured, domain-specific knowledge that can guide clinical decisions (Fig.1). Prior to any statistical analysis, data cleansing is usually performed which includes normalization, to ensure that different samples are comparable, and missing value handling to enable the use of methods that require all data points to be present. A plethora of imputation methods developed for microarray data [13] can be applied to proteomics as well [14]. Among these, methods with the underlying assumption that missing values result from protein expression that lies under the detection limit of modern mass spectrometers are frequently used. A typical task of clinical proteomics studies is to identify proteins that show differential expression between healthy and diseased states or between different subtypes of a disease. Although commonly established statistical methods, which achieve this task exist, distinguishing between expression differences due to technical variability, genetic heterogeneity, or even intra-sample variability and true disease-related changes require deep knowledge of statistical tools and good understanding of the underlying problems in the analysis of omics data.



**Fig. 1** Outline of a typical analysis workflow in Perseus. The workflow shows the process of converting data into information and knowledge. Statistical analysis can be used to guide the identification of biologically relevant hits and drive hypotheses generation. Various external databases, annotation sources, and multiple omics types can be loaded and matched within the software and together with powerful enrichment techniques allow for smooth data integration

For instance, testing thousands of proteins for differential expression is hampered by the multiple hypothesis-testing problem, which results in an increased probability of calling a protein a significant hit when there is no actual difference in expression. This necessitates the use of correction methods to increase the confidence of the

identified hits. The choice of the appropriate correction method depends on the balance between wrongly accepted hits (error type I) and wrongly rejected hits (error type II) that an experimentalist is willing to accept. For instance, permutation-based FDR [15] has a reduced error type II rate compared to the Benjamini-Hochberg correction [16]. Once the initial list of quantified proteins is narrowed down to only the significantly changing hits the question of their functional relevance arises. Enrichment analysis of protein annotations is the preferred method for deriving functional implications of sets of proteins and is applicable to both categorical (Fisher's exact test [17]) and expression/numerical data (1D enrichment test [18]). The outcome of such an analysis often offers a comprehensive view of the biological roles of the selected proteins through highlighting key pathways and cellular processes in which they are involved.

In this chapter, we provide a step-by-step workflow of bioinformatic analysis of proteomics data of luminal-type breast cancer progression. Commonly used analytical practices are described including data cleansing and preprocessing, exploratory analysis, statistical methods and guidelines, as well as functional enrichment techniques. All the steps are implemented as processes in Perseus [19], a comprehensive software for functional analysis of omics data.

- 1 The Methods section contains several modules covering the most frequently performed steps in the analysis of proteomics data. Often, a proteomics study benefits from a global overview of the data, which usually includes the total number of identified and quantified proteins, dynamic range, coverage of specific pathways, and groups of proteins. A good practice in data analysis is to start with exploratory statistics in order to check for biases in the data, undesirable outliers, and experiments with poor quality data and to make sure that all requirements for performing the subsequent statistical tests are met. Once the data are filtered and normalized appropriately, statistical and bioinformatic analyses are performed in order to identify proteins that are likely to be functionally-important. When the list of such proteins is small enough and direct links to the question of interest can be inferred using prior knowledge, follow-up experiments can be performed after this step to confirm the results of the statistical analysis. However, one of the advantages of mass spectrometry-based proteomics is the ability to unravel new discoveries in an unbiased way, for instance, through functional analysis. This analysis is often based on enrichment tests, which can highlight guiding biological processes and mechanisms.

### 3.1 Loading the Data

- 2 Go to the "Load" section in Perseus and click the "Generic matrix upload" button.
- 3 In the pop-up window, navigate to the file to be loaded.

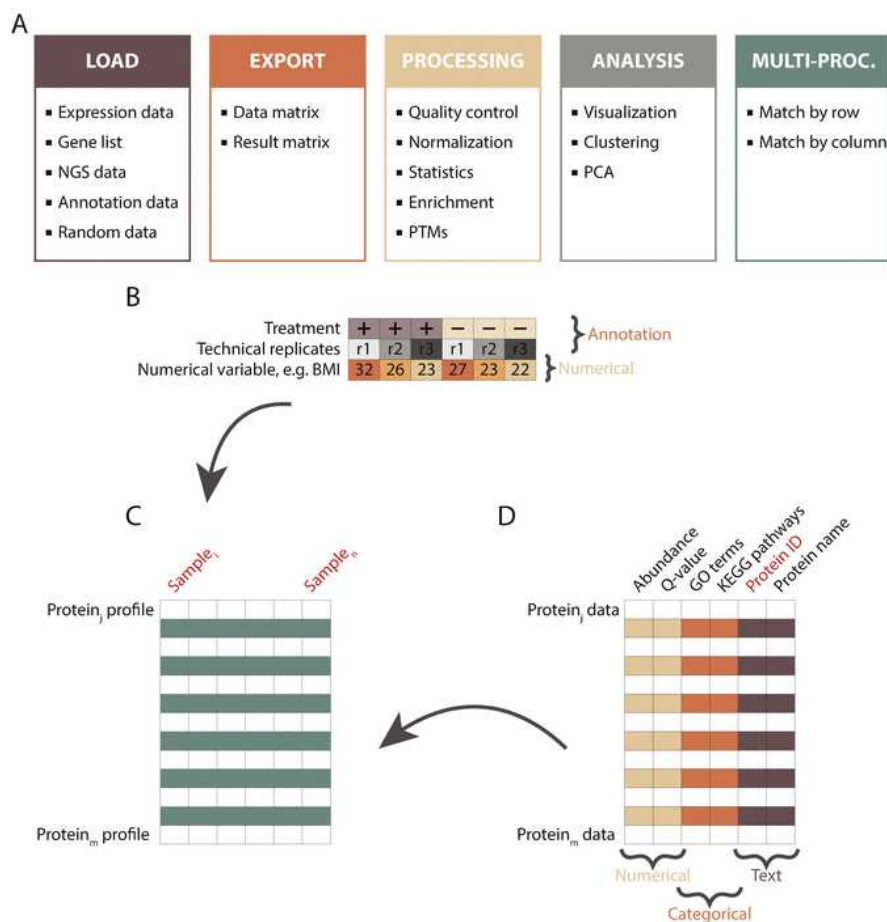
The input file format of Perseus is a tab-delimited txt file that contains a header row with the names of all columns. The type of data is specified during file loading. Make sure that the "Regional and Language Options" are set to English to avoid errors while reading numerical data such as decimal separators being wrongly interpreted.

- 4 Select all the expression columns and transfer them to the Main columns window. 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.

Different expression and meta data can be imported in Perseus and used for subsequent analysis. Common expression data are in one of the following formats: normalized intensities (e.g., LFQ intensity as described in [4], iBAQ as described in [22]) or ratios between heavy standard and light/non-labeled sample. Other data types that can be analyzed with Perseus are shown in Fig.2.

### 3.2 Summary Statistics

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



**Fig. 2**

Interfaces of Perseus and the augmented data matrix format. (a) Perseus extends over five interfaces, each of which includes various analysis and transformation functionalities and visualization possibilities. (b) Experimental design is specified as annotation (e.g., treatment vs. control groups) or numerical rows (e.g., variable concentration). Multiple annotation rows can be specified that allow biological and technical replicates to be analyzed together. (c) The data is organized in a matrix format where typically all samples are displayed as columns and all proteins as rows. (d) Additional protein information can be added in the form of Numerical, Categorical, or Text annotations

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.

- Go to "Processing → Filter rows → Filter rows based on categorical column" to exclude proteins identified by site, matching to the reverse database or contaminants.

Reverse, identified by site and contaminant proteins have to be marked in a categorical column before these filters can be applied. These are automatically set when MaxQuant output tables are used for analysis in Perseus. Additional filtering options can be used to remove proteins based on a quantitative measure such as a minimum number of quantified peptides or a maximum *q*-value.



- 7 Transform the data to a logarithmic scale by going to “Processing → Basic → Transform” and specifying the transformation function (e.g.,  $\log_2(x)$ ).
- 8 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.

### 3.3 Filtering

- 9 Use the workflow window to select the *InitialData* matrix data by clicking on it (see **Note 5**).

Different activities have different output results including a data matrix with the same expression values and additional columns containing the results of the analysis or a new data matrix containing only the output of an analysis activity. An activity is always performed on the data matrix and specific tab for that matrix that is active at the moment.

- 10 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.

Depending on the nature of missing values, different filtering strategies may be employed and are supported in Perseus. For example, if large differences between groups are expected with proteins having very low expression level in one of the groups, filtering based on a minimum number of valid values in at least one group would be a more suitable approach than filtering for a minimum number of valid values in the complete matrix.

### 3.4 Exploratory Analysis

- 11 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*.
- 12 Explore the visualization options in the Histogram panel by testing the functionality of each of the buttons (e.g., *Properties*, *Fit width*, *Fit height*).
- 13 Click on the *pdf* button to export the plot.

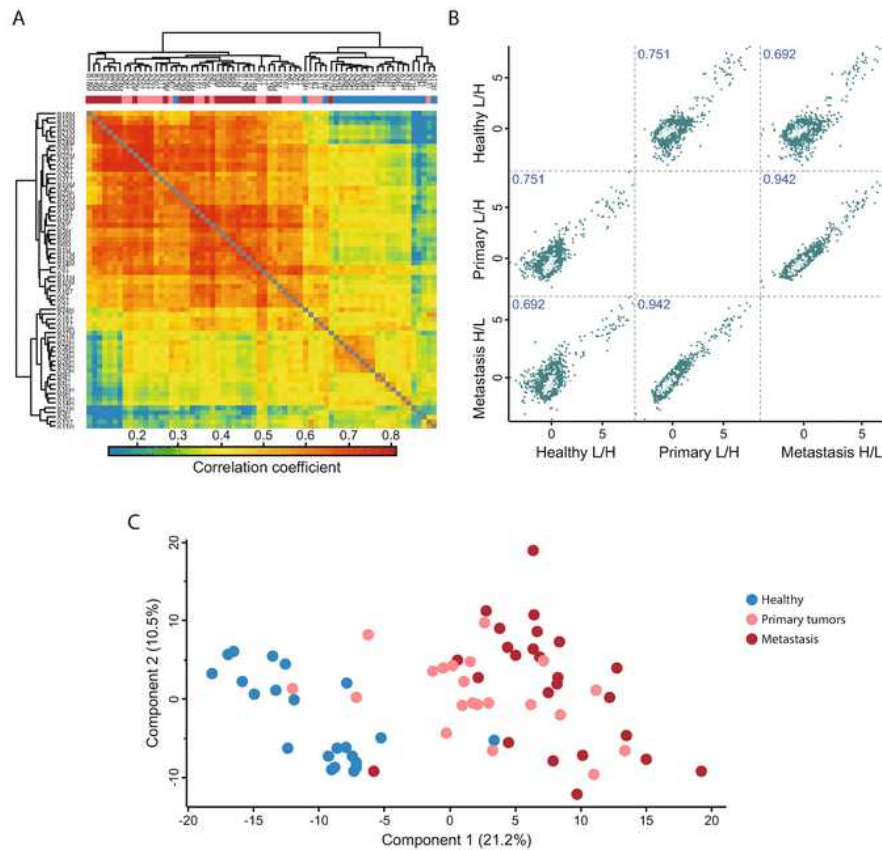
All the plots can be exported in figure-ready formats such as pdf, tiff, or png.

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

- 16 Adjust the plot using the *Fit width* and *Fit height* options and resizing the plot window.
  - 17 In the drop-down menu "Display in plots" in the plot window, select *Pearson correlation*.
  - 18 Select a scatter plot by clicking on it. The selected plot will be shown in an enlarged view.
  - 19 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.
  - 20 Switch back to the "Data" tab to continue with the analysis.
  - 21 "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.
  - 22 To visualize the sample correlations, go to "Analysis → Clustering/PCA → Hierarchical clustering." Use the *Change color gradient* to set a continuous gradient similar to the one in Fig. 3a.
  - 23 Export the plot by clicking on the *pdf* button.
  - 24 Navigate back to the previous data matrix by clicking on it in the workflow panel.
  - 25 Principal component analysis requires all values to be valid. To remove all protein groups with missing values, repeat Section 3.3, **step 2** setting the percentage parameter to 100.
- Very stringent filtering is usually not recommended, as a large amount of the data will be lost. Instead milder filtering combined with imputation may be more appropriate.
- 26 Go to "Analysis → Clustering/PCA → Principal component analysis" and click *ok*. Explore the sample separation (dot plot in the upper panel) and the corresponding loadings (dot plot in the lower panel).
  - 27 In the table on the right of the PCA plot, select a set of samples (e.g., all samples that belong to one experimental condition) and change their color by clicking on the *Symbol color* button and selecting the desired color.



- 28 Check the contribution of other components by substituting Component 1 and 2 with other components from the drop-down menu. Find the components that show sample separation according to the experimental conditions (see Fig. 3c).
- 29 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.



**Fig. 3** Exploratory analysis outputs in Perseus. (a) Hierarchical clustering of all the samples based on the correlation coefficients between them reveals higher similarity between primary and metastatic tumors versus healthy tissue samples. (b) Multi-scatter plot of averaged profiles among the three main groups clearly represents the disease progression by highlighting strong similarities between subsequent stages, e.g., healthy tissue samples are more similar to primary tumors than to metastasis (correlation coefficient 0.76 vs 0.69), whereas primary tumors are most similar to metastasis ( $R = 0.94$ ). The category Cell division is highlighted in bright green in all pairwise comparison plots. (c) Principal component analysis (PCA) attributes the largest variance to the difference between healthy (blue dots) and cancer tissues (pink and red dots) (Component 1, 21.1%) and shows that primary and metastatic tumors (pink and red dots respectively) are difficult to distinguish

### 3.5 Normalization

- 30 Navigate back to the data matrix before filtering for 100% valid values (Section 3.3, **step 2**).
- 31 Go to "Processing → Normalization → Z-score." Change the *Matrix access* parameter to *Columns* and select the *Use median* option. In the new data table, plot histograms for the same subset of samples as in Section 3.4, **step 1**.

Data normalization is not always necessary. Different types of normalization can be applied on the data to correct for systematic shifts or skewness and to make samples comparable.

### 3.6 Experimental Design

- 32 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.

Regular expressions can be used to derive the experimental design from the sample names (“Action → Create from experiment name”). Additionally, a template txt file can be written out, edited in an external editor program, and read in to indicate the experimental design.

### 3.7 Loading Annotations

- 33 Go to the drop-down menu indicated with a white arrow at the top left corner of Perseus and select “Annotation download.”
- 34 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).
- 35 Download the file to the Perseus/conf/annotations folder.
- 36 Go to “Processing → Annot. columns → Add annotation.” Select the file from the previous step as a Source.
- 37 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).
- 38 Select several categories of interest to be overlaid with the main matrix and move them to the right-hand side. Click ok.

### 3.8 Differential Expression Analysis

- 39 Go to “Processing → Tests.” From the menu select the appropriate test. For the data set used in this chapter, the *Multiple-sample tests* option should be chosen, as there are more than two conditions that are compared. The default parameters do not have to be changed.

Analysis of differentially expressed proteins depends on the number of compared conditions, the underlying distribution properties, and the availability of biological replicates. For example, data sets with one condition with replicates should be analyzed with *One-sample tests*, with two conditions—with *Two-sample tests*, and with more than two conditions—*Multiple-sample tests*. Paired samples test and tests abolishing the requirement for equal variance are also available.

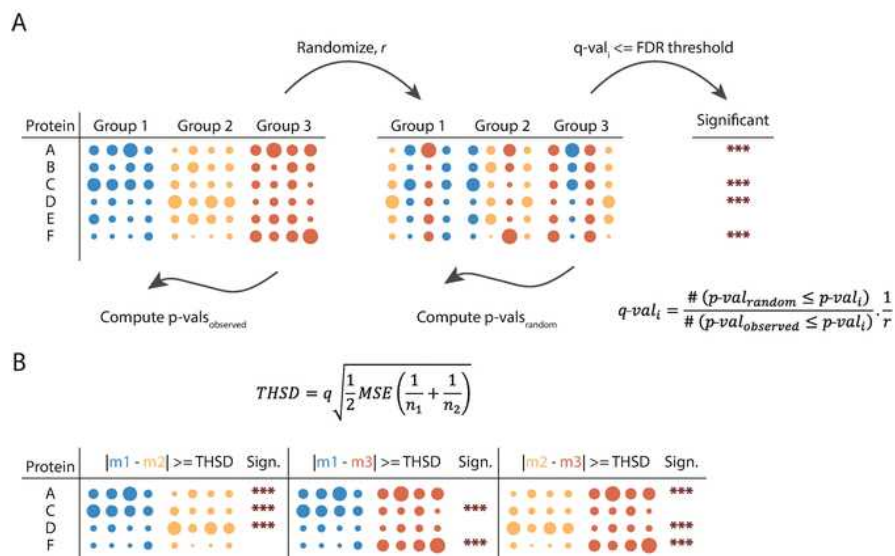
- 40 Specify the categorical row that contains information about the experimental conditions of the samples that will be used in the differential analysis in the *Grouping* parameter.
- 41 Keep the default value of 0 for the *S0* parameter, to use the standard t-test statistic. Change the parameter to use the modified test statistic approach described by Tusher et al. [15].
- 42 Select the multiple hypothesis testing correction method to be used by specifying the *Use for truncation* parameter (see Note below and Fig.4a).

The method with largest power *Permutation-based FDR* is recommended and at least 250 repetitions are suggested. In case of technical replicates, these have to be specified as a separate grouping (see Section 3.7, **step 1**) and selected with the "Preserve grouping in randomizations" option. Failure to specify technical replicates will result in wrong FDR calculation. The more conservative Benjamini-Hochberg correction can also be used when a lower number of Type I errors at the cost of lower sensitivity are desired.

- 43 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.
- 44 Inspect the output table. It contains three new columns: *ANOVA significant*, *-Log ANOVA p-value*, and *ANOVA q-value*.

The "Significant" column contains a "+" if a protein met the selected significance threshold (usually *q*-value). Additionally, *p*-values (probability of type I error) and the corresponding *q*-values (corrected *p*-value) are provided in the output table.

- 45 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.
- 46 Go to "Processing → Tests → Post-hoc tests." Set the *Grouping* parameter to the same grouping that was used for the ANOVA test (see Section 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** below and Fig.4b).



**Fig. 4**

Differential expression and multiple hypothesis testing. (a) Multiple hypothesis testing correction using a permutation-based false discovery rate approach is shown. Labels are randomly swapped between the three groups (blue, yellow, and red). The Randomization is repeated  $r$  times. ANOVA  $p$ -values are computed both on the measured and the permuted data and local FDR values ( $q$ -values) are computed as the fraction of accepted hits from the permuted data over accepted hits from the measured data normalized by the total number of randomizations  $r$ . All hits with a  $q$ -value smaller than a threshold are considered significant. (b) To determine the exact pairwise differences of protein expression Tukey's Honest Significant Difference (THSD) test is used on the ANOVA significant hits. If the mean difference between two groups is greater than or equal to the corresponding THSD, the difference is considered significant between the compared groups.  $q$ : constant depending on the number of treatments and the degrees of freedom that can be found in a Studentized range  $q$  table; MSE: mean squared error;  $n_1, n_2$ , number of data points in each group

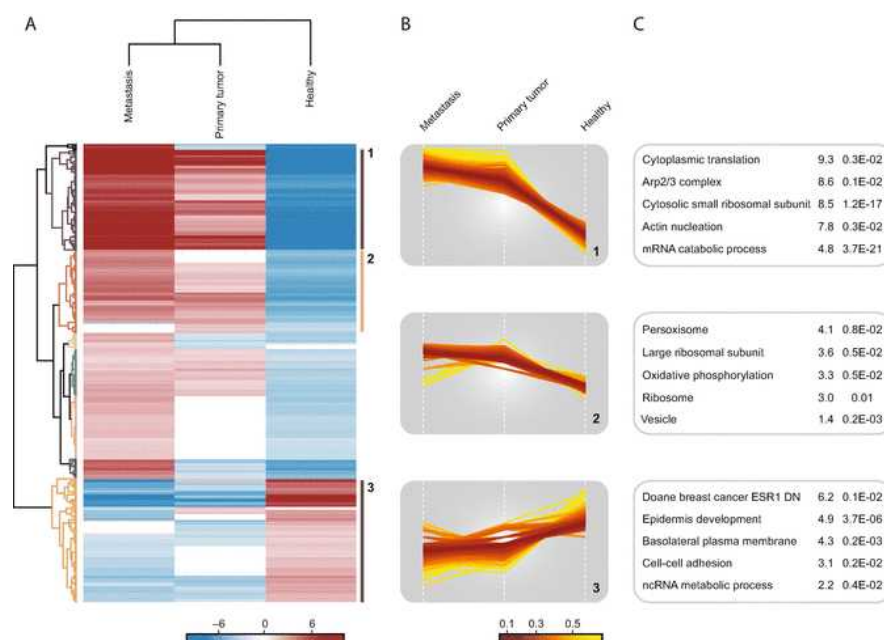
Tukey's honestly significant difference (THSD) is a post-hoc test that when performed on ANOVA significant hits determines in exactly which pairwise group comparisons a given protein is differentially expressed.

### 3.9 Clustering and Profile Plots

- 47 Go to "Analysis → Clustering/PCA → Hierarchical clustering." Keep the default parameters and click *ok*.
- 48 Inspect the resulting heatmap and the relationship between the groups and the proteins.
- 49 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.
- 50 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.

Clusters can be defined by clicking on the respective nodes in the protein tree or based on the precise distance measure used to compute the tree. To use the latter option, click on the "Define row clusters" button and specify the desired number of clusters, which will then be automatically defined.

- 51 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).
- 52 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.



**Fig. 5** Enrichment analysis highlighting important pathways and processes. (a) Hierarchical clustering of proteins found to have differential expression between pairs of disease states. High and low expression are shown in red and blue respectively. Various clusters of protein groups are highlighted in the dendrogram. (b) Profile plots of three selected clusters showing distinct behavior with respect to the three disease states are shown: 1 strongly increased expression in tumor tissues; 2 moderate increase in tumor tissue; and 3 decreased expression in tumor samples. (c) Functional analysis of protein annotation terms resulted in multiple categories that were enriched in the three selected clusters. The enriched terms and the corresponding enrichment factor and *p*-value are shown

### 3.10 Functional Analysis

- 53 Go to "Multi-proc. → Matching rows by name." Both *Base* and *Other* matrices point to the last matrix.
- 54 Click on *Base matrix* and then in the workflow window select the data matrix that was generated before filtering for ANOVA significant hits (Section 3.9, **step 6**).
- 55 In the pop-up window set *Matching column in matrix 1 and 2* to a common identifier (e.g., *Protein IDs*).
- 56 In the categorical columns section, transfer the category *Cluster* to the right hand-side. Click *ok*.

The matching step is necessary in order to define the correct background against which enrichment will be computed. Too small (only significant hits) or too large (the complete proteome, even if not detected with MS

analysis) introduces bias in the enrichment results.

- 57 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 Fisher’s exact test and multiple hypotheses correction.

The enrichment output table contains information about the values used to compute the contingency table for the Fisher’s exact test (e.g., category and intersection size), the enrichment factor, and the statistical significance of the hit indicated by *a*-*p*-value and the associated false discovery rate.

In summary, this chapter provides a complete protocol for fundamental analysis of proteomic data, starting from data upload and transformation and ending with identification of proteins, characteristic of the specific disease progression stage, and the underlying processes in which they are involved. The described analytical methods and visualization tools are integrated in Perseus, a freely available platform for analysis of omics data, which provides a comprehensive portfolio of analysis tools with a user-friendly interface [19]. A special emphasis is placed on employing statistically sound methods in the analysis of large data, avoiding wrong interpretation and extracting maximum information. More advanced computational techniques such as supervised learning are also supported and are often instrumental for the analysis of complex data where genetic and intra-tumor variability may pose challenges. Moreover, Perseus is being continuously developed to integrate analysis of various data types, including posttranslation modifications, sequence information, as well as to allow deeper functional interpretation through network and pathways analysis.