Design and implementation of analysis pipeline for single cell type proteomics data

By

Lukas Gamp

in partial fullfilment of the requirement
for the degree of MSc
in Bioinformatics

mm yy

Abstract

(the spacing is set to 1.5)

no more than 250 words for the abstract

- a description of the research question/knowledge gap – what we know and what we don't know
- how your research has attempted to fill this gap
- a brief description of the methods
- brief results
- key conclusions that put the research into a larger context

Contents

Acknowledgements

Thank you for following this tutorial!

I hope you'll find it useful to write a very professional dissertation.

1 Introduction

1.1 Proteomics

The proteome is referred to the sum of all proteins of a given sample at a given time. In the past several quantitative and qualitative assays were used to enlighten the protein composition of a sample.

An early approach of qualitative analysis of the cellular proteome involved labeling with fluorescent antibodies and imaging. The major disadvantage of this technique was the limitation to only stain a few proteins per cell. For quantification procedures such as single-cell Western blots, immunoassays or CyTOF have been used. Other disadvantages are the ability to permeate cells, accessibility and binding of the epitope and the creation of specific antibodies for a given protein (Budnik et al. 2018).

One of those techniques involved RNA-sequencing. Since RNA involves also non-coding RNA, the amount of RNA is often not proportional to the content of proteins in a cell. So the proteinaceous content of a cell was only predicted and quantitative analysis was not possible.

1.2 Mass Spectrometry

Mass spectrometry enables qualitative and quantitative analysis of the entire repertoire of a biological sample. The availability of gene sequences in databases and the ability to match proteins against those sequences with computational methods makes it possible to identify alterations of a sample on a protein level. These alterations can rely on the sequence level or could be to post-translational modifications (PTMs) such as phosphorylation, methylation or else (Aebersold & Mann 2003).

Mass to charge ratio (m/z)

2 Materials and Methods

2.1 Materials

For analysis two types of cells were used. One type is the Jurkat-based cell line (J-lat) with integrated HIV.

The other type of cells are macrophages with a sample size of 72 cells. The analysis is done with two groups. A HIV negative (HIV-) control group and a HIV positive (HIV+) group.

2.1.1 Cell Isolation

2.1.2 Lysis

2.1.3 Digestion

2.1.4 Labeling techniques

For differential analysis proteins need to be labeled to compare mass to charge intensities in order to quantify observed peptides. Since mass spectrometry is not a quantitative technique by itself, the peak height or area does not reflect the abundance of a peptide. Physicochemical properties of the proteins can change the ionization efficiency and detectability of the target. However, when comparing the same analyte between multiple runs of labeled proteins, differences in the mass spectrum reflect the abundance of those. Labels should be chosen to change solely the mass of the sample and to not affect folding or other inherent properties of the protein.

2.1.4.1 Metabolic labeling Feeding cells with aminoacids containing heavy isotopes, is the method of choice in order to label peptides at the earliest possible level. This atoms can be heavy nitrogen in aminoacids or salts in fertilizer for plants. Mass shifts are proportional to the isotopes incorporated during biomass production and are visible after proteolytic cleavage. Stable isotope labeling in cell culture (SILAC) was presented in the early 2000s. This method used heavy aminoacid enriched media to feed cells, in order to quantitatively analyze expression profiles.

2.1.4.2 Isobaric labeling

2.1.4.2.1 Tandem mass tag (TMT) Tandem mass tag (TMT) reagents enable to differentiate multiple samples analyzing in one MS run. The samples are labeled individually and pooled afterwards, this procedure is called multiplexing. TMTs have the same charge and differ only by their isotopic masses, the peaks found for each sample are called reporter ions (RI). Each RI and sample is interpreted as one channel in downstream analysis. The identification of these RI leads to an enrichment and identification of low abundance peptide ions which is common especially in single-cell techniques. With this technique it is possible to quantify proteins and differ low abundant proteins from background noise. The disadvantage of isobaric labeling is, that the co-fragmentation signals can be observed in the spectrogram and the data needs to be normalized in order to remove unwanted contribution (Marx 2019, Budnik et al. 2018). Furthermore TMTs have an isotopic distribution according to the distribution found in nature. This can be corrected during data-acquisition as a defined spread in other channels.

2.1.5 Instrumentation

2.1.5.1 Liquid chromatography In order to separate proteins according to their chemical properties, size or species a liquid chromatography (LC) is recommended before ionization.

2.1.5.2 Mass Spectrometry

2.1.5.2.1 Ionization In order to analyze a biological sample consisting of proteins in solution the liquid needs to be vaporized into gas phase. Two techniques are capable of this procedure. Electrospray ionization (ESI) pushes the analyte through a capillary and applies an electric current to the liquid, vaporizing the sample to a charged aerosol. Biomolecules are fragmented according to their chemical properties and can be further handled in the mass spectrometer. The fragmented biomolecules are now in charged droplets separated by their charge on the surface, splitting further into smaller droplets until they become a gas phase ion. Two physical models describe the process from gas phase to ion called “The ion evaporation model” (IEM) and “The charge residue model” (CRM). In the ion evaporation model (by Iribarne and Thomoson) the droplets shrink by evaporation until ions are expelled. The model had its limitation by explaining same evaporation rate constant among ions with different chemical properties. In the charge residue model the assumption of one molecule per droplet leads to an ionization rate constant, which is independent of the ion itself and relies solely on the generation of the droplet and the efficiency of the solvent (Wilm 2011).

Matrix-assisted laser desorption/ionization (MALDI)

2.1.5.2.2 MS.1

2.1.5.2.3 Coupled mass-spectrometry (MS/MS) & MS.2 In order to enhance sequence identification, two MS devices are built in series. In the first run (MS1) the m/z is determined and the molecules are passed to the next device. Upon passing the molecules are fragmented into smaller ions and analyzed by the second MS. The fragmentation highly depends on the chemical bonds found in the molecule. The majority of these breaks occur on the peptide bond of the protein, although this is not guaranteed for all bonds and so it can happen that certain peptide ions have a low abundance (Budnik et al. 2018). These low abundant peptides will not be detected, hence the problem needs to be faced with another approach . A solution for this problem is molecular barcoding with labeling mentioned in the chapter labeling.

2.2 Data

2.2.1 Acquisition

Acquisition of the data was done with MaxQuant (Cox & Mann 2008) software package.

2.2.1.1 MS-Spectrum Each peptide is reflected by its‘ indivual fingerprint in the ms-spectrum. The fingerprint is based on the chemical properties and modifications of aminoacids. These aminoacids can be calculated through their m/z ration and after that interpreted as an aminoacid sequence. Due to fragmentation of the protein only peptide sequences are visible in the spectrum. In order to identify proteins, peptides are matched against a sequence database (Cox & Mann 2008). Sequence Databases are simple .fasta files, which can be downloaded on the uniprot webpage (www.uniprot.org).

Since ms data has a high resolution, algorithms are used to convert the raw signal to an interpretable form. MaxQuant is one of many sofware packages to process the data and provides it for further analysis and statistical testing. Other software solutions are Protein Discoverer Thermo Fisher or even packages for R. In this publication we will mainly focus on the data-acquisition with MaxQuant (Cox & Mann 2008).

2.2.1.2 Three-dimensional peak detection The three dimensions of the data are: m/z ratio, intensity and retention time. The algorithm finds local minima of the function in order to seperate peaks from each other. The centroid of the peak is detected by fitting a so called gaussian peak shape

fitting. This can be interpreted as finding the peaks of each m/z spectrum as a function of time.. The centroid of the peak refers to an isotope.

2.2.1.3 Deisotoping To decrypt the istopic distribution of a biomolecule, MaxQuant creates a vertex of every single peak and connects them with their possible isotopic counterparts by finding the proportion of mass of an average aminoacid to its‘ respective isotope (averagine (Senko et al. 1995)). Isotoping is the term of such procedure and it is enabled with graph theory. After this procedure the amount of data points are reduced by a tenfold and a single peak reflects a small biomolecule.

2.2.1.4 Label detection The next step in data-acquisition is the detection of labels for quantification. Isotopic pairs of the label (e.g. N13, N14, N15) contained in the tag or aminoacid can be identified by convoluting the two measured isotope patterns with the theoretical isotope patterns. With a least-square method the best fit is found iteratively and the channel/sample can be identified.

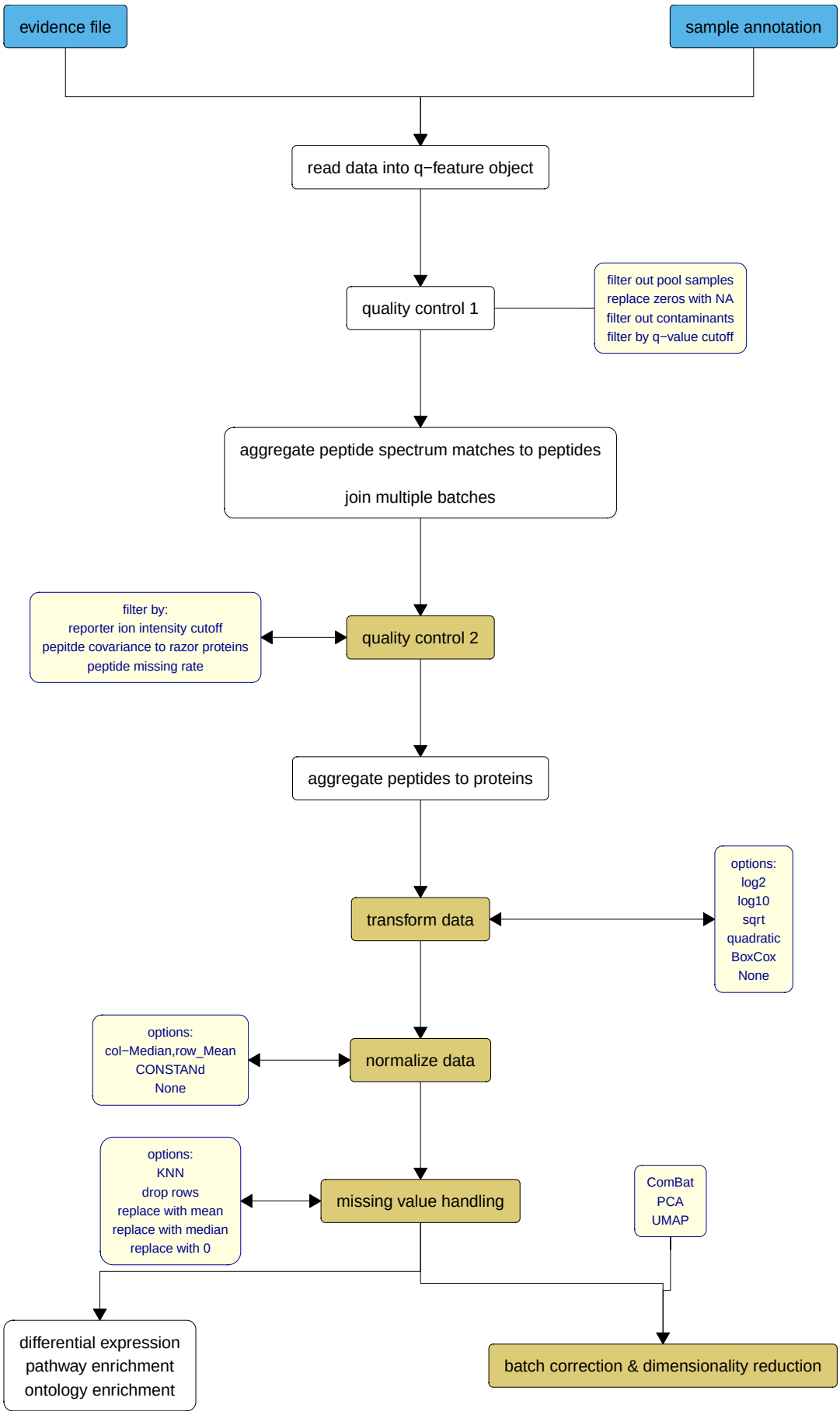
2.2.1.5 Improving peptide mass accuracy The intensity-weighted average of the ms peak centroids (as described in the 3D peak identification) refers to the mass of the peptide. Corrections highly depend on the analyzer, MaxQuant uses for an orbitrap typed analyzer a correction value of 1ppm. Autocorrelation between centroids is compensated by only using well-identified peptides. As published the mass precision within a ms experiment ranges around 10e-7.

2.2.1.6 Peptide search Biomolecules can now be searched in a database in forward and reverse direction. The peptide identification (P-) score indicates the fit of the data to the found sequence in the database according to the length of the peptide and is used to calculate the posterior error probability (FDR). The calculation of the false discovery rate is then calculated by taking FDR into contrast.

2.2.1.7 Protein assembly After these calculation the identified peptides can be aggregated according to it‘s respective protein and quantified. The mentioned metrics indicating the performance of the peptide search can be used in downstream analysis. A so called razor peptide indicates the group with the highest number of identified peptides. Quantification is enabled by taking only unique peptides into contrast. Posterior error probabilities, which refers to the chance that the found peptide is a random event, are multiplied and only distinct sequences with the highest-scoring are used.

2.2.2 Data processing

Further analysis is done with R and respective packages such as bioconductor. Since there is no state of the art established, analysis varies upon experimental design. The workflow of the analysis will be processed in a so called pipeline streamlining the data through steps where individual results can be observed in visualizations and indivdual calculations will be adapted according to user demands and experimental properties.


evidence file
sample annotation
read data into q−feature object
quality control 1
filter out pool samples
replace zeros with NA
filter out contaminants
filter by q−value cutoff
aggregate peptide spectrum matches to peptides
join multiple batches
filter by:
reporter ion intensity cutoff
pepitde covariance to razor proteins
peptide missing rate
quality control 2
aggregate peptides to proteins
options:
log2
log10
sqrt
quadratic
BoxCox
None
transform data
options:
col−Median,row_Mean
CONSTANd
None
normalize data
options:
KNN
drop rows
replace with mean
replace with median
replace with 0
missing value handling
ComBat
PCA
UMAP
differential expression
pathway enrichment
ontology enrichment
batch correction & dimensionality reduction

2.2.2.1 Reading the data After processing MaxQuant creates a directory containing all results as .txt file. The evidence.txt file include all peptide to spectrum matches (PSM) with their respective proteins and statistical parameters.

Example fo basic parameters and derivations include:

- Peptide sequence
- Mass to charge ratio (m/z) for all scans (eg. MS1, MS2)
  - Mass
- Retention time
- Precursor Ion Fragment
  - source of the detected ion also referred as mother ion
- Fraction of total spectrum
- Base peak fraction
- Reporter intensity (RI)
  - corrected RI
- Posterior error probability (PEP)

2.2.2.2 Object oriented programming In order to streamline the analysis of multiple experiments, object oriented programming can be applied. The approach in R is to create a so called Q-feature object, which contains all variables and metadata in a hierarchical structure. The structure enables sub setting for further analysis (Vanderaa & Gatto 2021).

2.2.2.3 Zero values Peptides with low abundance are often set to zero during analysis. However, assigning a value of zero may incorrectly suggest that the sample does not contain the respective peptide. Given that it is highly unlikely for a biological cell of a comparable type and function to not contain a particular protein, replacing the zero value with "not applicable" (NA) is crucial for understanding and interpreting MS data.

2.2.2.4 Exclude reverse matches/contaminants Peptide sequences matching to the reverse protein sequences (=decoy database) are considered as possible contaminants. These matches can be excluded from further analysis.

2.2.2.5 Filter according to precursor ion fraction (PIF) During mass spectrometry, the ions detected in MS1 are further fragmented through collision during multiple MS runs. The resulting product ions are derived from precursor ions (also known as mother ions or parental ions). Contaminant peptides can co-migrate in this process and can be distinguished by the lower fraction of their respective precursor ions (Tannous et al. 2020). These peptides need to be filtered out during the analysis pipeline. A cutoff value, referenced in the SCoPE2 pipeline (Specht et al. 2021), is applied in the user interface, but it can be adjusted according to the needs of the biologist.

2.2.2.6 Filter by q-value The next step for quality control is the exclusion of samples with a high false discovery rate (FDR). When applying multiple statistical testing (e.g. t-Test) the obtained p-values can be considered as biased, because the probability to observe a significant will iteratively increase with each test performed. Corrections in statistics are an approach to compensate for the multiplicity of testing. There are many ways to do this compensation like the Bonferroni method or Benjamini-Hochberg‘s FDR. In Mass spectrometry the common “way to go” is calculating a false discovery rate, by dividing false PSMs (=hit of the decoy database) through the total number of PSMs above the peptide-spectrum matching score. The peptides spectrum matching score is defined as -10log10(p). Whereas the p-value is defined that the hit is done by chance. The calculation of the score is highly dependent on the data acquisition method used. MaxQuant uses Andromeda , an integrated search engine. Proteome Discoverer from Thermo Fisher utilizes different engines such as Mascot or Minora. As published by J.Cox in 2011 Mascot and Andromeda showed similar performance when comparing FDR values as a function of coverage. However the observed performance can be lower when dealing with a decreased coverage (Cox et al. 2011).The threshold for accepting an FDR of an individual PSM is described as q-value.

2.2.2.7 Peptide spectrum match (PSM) aggregation to peptides In data science, aggregation refers to a row-wise operation that merges data based on a particular column using a specific function. In the context of processing from peptides to spectrum matches, the desired column is the peptide sequence. To account for different distributions across multiple assays, the median of the channel is used as the function to aggregate multiple matches into one.

2.2.2.8 Join assays when observing multiple comparable batches at once Sample size is often a limiting factor in hypothesis testing. A strict quality control and the fact that TMT reagents are only available up to 18-plex can reduce the number of observed samples below the critical

threshold, leading to an early end of analyses. To overcome this limitation, the provided software is capable of processing multiple runs simultaneously, allowing for testing of multiple batches and increasing the number of samples that can be included in the analysis.

2.2.2.9 Calculate reporter ion intensity (RI) and filter according to median RI

Columns which do not meet the desired intensity can be filtered by a threshold set on the RI. The median RI can also be used to check if an entire channel has a lower detection level. This can be due two reasons. One is the expression level of the given proteins in a cell. Meaning, that the expression of the observed cell type is simply lower than the other type. Another one could be a spillage of TMT detection in other channels due incorrect or missing correction of the TMT isotopes.

2.2.2.10 Calculate and filter according to median coefficient of variation (CV) per cell/channel

Depending having a bulk sample or single-cell sample choosing a minimum of observed peptides and a cutoff value for the CV, changes the level of confidence in the peptide data. The coefficient of variation of a peptide is considered as the ratio of the standard deviation to the mean and describes the relationship of the observed peptide signal over multiple proteins (=razor proteins). Peptides having a high coefficient of variation over many razor proteins are considered as noise and need to be filtered out before statistical analysis.

2.2.2.11 Remove peptides with high missing rate

Although missing value imputation can be performed during the analysis of multiple batches, peptides with missing detections across channels can be problematic for quantification. The proteomic composition of a biological sample is similar between replicates and even across groups. However, the threshold of missingness (described as a fraction of the row) can be set in the user interface and adjusted to enable different experimental designs.

2.2.2.12 Aggregation of peptides to proteins

Similar to the already explained previous aggregation step the peptides will be further processed into their respective proteins after the quality control on the peptide level is performed. Finally an expression matrix for every protein and their respective channel column is returned. Each channel refers to a sample and reflects contains the intensities of each protein found in the biological specimen.

2.2.2.13 Transformation of protein expression data Data transformation applies a function to each value of a matrix or array, so that:

$$y_i = f(x_i)$$

Depending on the distribution of the values in the observed expression set, different transformations can be applied to fulfill assumptions for statistical testing. In the shiny application, various procedures are implemented and can be further expanded upon request from the user. The macrophage analysis done by Specht et al. mentioned in their SCoPE2 publication (Specht et al. 2021) uses the logarithm to the base 2 to spread a compacted distribution and remove skewness in the dataset. This transformation was used as a reference when comparing methods. However, using the logarithm to the base 10 may be an easily interpretable way of defining expression data and will also be facilitated by other transformation methods such as the boxcox method, which is also implemented in the application. When observing a wide distribution of small and large values of the expression set in a histogram, a square root transformation can help increase the variability of smaller values and decrease the variability of larger values. Practically impossible, but in data-driven science, occasionally negative expression values may be present, and they can be converted into positive values by taking each one to the power of two. Depending on the distribution, the user has to decide which transformation to consider and verify that statistical assumptions are met after application with visualizations such as histograms, qqplots, and MA-plots. For statistically inexperienced users, the boxcox transformation can help with this decision. Developed in 1964 by Box and Cox (Sakia 1992), this method applies a linear model against 1 on the data to determine the statistical parameter lambda by the maximum likelihood method. The log-likelihood method takes the 95% confidence interval for the parameter lambda and the final lambda is chosen as the value with the highest log-likelihood value.

λ	Transformation
-2	$\frac{1}{x^2}$
-1	$\frac{1}{x}$
-0.5	$\frac{1}{\sqrt{x}}$
0	$\log x$
0.5	$\sqrt{x}$
1	x
2	x^2

Depending on the size of lambda, a certain function will be automatically applied to each value of the expression set in order to introduce normality. In terms of usability, the boxcox method is the most efficient transformation method, taking the decision of which calculation to apply away from the user.

2.2.2.14 Normalization The term normalization is ambiguous in data-science and will be explained briefly in this chapter. Starting with converting the data to the Z-distribution (Gaussian result), where 0 represents the mean and 1 represents the standard deviation, in vectorization every feature (or protein) is represented as a vector that points in a specific direction in the unit sphere, normalization can have different meanings. Normalization is considered as part of the scaling procedure and ensures that every feature contributes equally to our statistical model and hinders large values from biasing the model in a particular direction. However, this could also have negative consequences for the modeling procedure, as it may reduce the impact of important features on the dataset. Depending on the algorithms later used in the data processing, the method of choice for normalization can change the results significantly. The K-means clustering algorithm, for instance relies on the distances between distinct data points by minimizing variance of the squared euclidean distances. This method will be applied in the missing value imputation which will be the next step applied to our data. Therefore, we want to ensure that we obtain the most accurate imputed values possible. The first normalization method available in the user interface is column-wise median and row-wise mean normalization. This method was chosen by Specht et al in the SCoPE2 publication 2021 (Specht et al. 2021) and worked as a reference during the development process to benchmark other methods. When dividing each value of an expression set by the median of the particular column, the procedure is considered as a column-wise median normalization. Although the mean-normalization works the same way but takes the mean as the second variable. In the analysis pipeline both procedures were applied on the dataset. However when the logarithm as the transformation method was chosen as transformation method, the pipeline automatically switches from a division to a subtraction, since: $\log_a\left(\frac{u}{v}\right) = \log_a u - \log_a v$. After applying this normalization method, each value can be interpreted as a distance or fraction of the mean or median of the corresponding column or row, enabling fair comparisons between features and samples. The second method for normalization included is the CONSTANd method, which was proven suitable for relative quantification in the field of proteomics (Maes et al. 2016, Houtven et al. 2021), but can also be applied on RNAseq data. CONSTANd uses a technique called matrix raking, which employs the RAS algorithm. The expression set reflects the nonnegative real (m, n) matrix A were the bi proportional constrained matrix problem will be solved by finding the (m, n) matrix B which equals

to $diag(x) * A * diag(y)$. Whereas $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. The solution to this problem involves finding the row sum of $B = u_i$ and the column sum of $B = v_j$, where i and j denote the row and column indices of the matrix, respectively (Bacharach 1965). In other words, the matrix will be alternatively manipulated on rows and columns until the mean of both equals 1. In order to apply the method and yield true results, one has to consider 3 major assumptions between sample types. This assumption can be observed in the MA-plot which indicates the differences of two samples. If we have the intensities of two samples, R and G, we can plot them on a graph with the x-axis as $M = \log_2(\frac{R}{G})$ and the y-axis as $A = \frac{1}{2}\log_2(RG)$. A comparison between all sample types of identically processed sets is needed before calculation and the assumptions must be fulfilled.

- **1 The majority of proteins are not differentially expressed**
  Density of the dots decreases when looking from the middle of the cloud towards the edges.
- **2 Up- and downregulation is balanced around the mean expression**
  The scatterplot of the data points is symmetrical around the mean on the horizontal axis.
- **3 Systematic bias correlates with the magnitude of expression**
  The symmetry axis is approximately horizontal.

If the MA-plot does not visualize any of these violations the CONSTANd method can be applied on the dataset and the biological meaning of the sample types can be assessed. The method is utilized by the bioconductor package handler. However if the dataset already consists of balanced features and samples or the user wants to benchmark the methods applied, normalization can be skipped also.

2.2.2.15 Missing value handling When analyzing multiple batches simultaneously, it is possible to encounter situations where a particular protein is not detected in one of the batches. From a biological perspective, it is highly unlikely that a cell completely lacks a single protein, as it would imply a complete loss of function for that protein. Therefore, an intensity value of 0 is close to impossible for most proteins. In order to handle this dark space in the expression set, several methods can be chosen in the application. A variety of function employs replacing the missing values with the mean or median of the rest of the matrix. This procedure introduces no bias in the data if the missingness is low, although differential expression for the sample carrying the missing feature can not be expected as well. Even a more conservative approach is dropping the rows for the feature with the missing value in one of the observed samples. A common practice in proteomics is utilizing the K-nearest neighbor algorithm (=KNN), which imputes the missing values by a euclidean metrics

of the neighboring values in the columns where the feature is not missing. K denotes for the count of neighboring values of a gene and can be selected upon preference. The euclidean distance between two points is defined as $d = \sqrt{a^2 * b^2}$, the K nearest neighbors average value will be assigned to the intensity of the missing feature. Since the average is a statistical metric which can be biased easily by skewed distributions, normalization is an important task before performing missing value imputation. KNN was benchmarked for the regression of protein abundance intensities and showed an Area under the Curve (AUC) above 0.7 when testing on a human proteome test set (Lan et al. 2013).

2.2.2.16 Batch correction Since mass spectrometry experiments are highly influenced by several factors, such as sample processing, the technician, manufacturer and production of the device, the reagents etc., working in multiple batches introduces further variance in the protein data. This variance would overshadow the biological differences across sample types when performing dimensionality reduction. A common misconception in biology is, that batches need to be corrected before differential expression analysis, which is the final aim of the developed pipeline. It appears convenient to remove the batch effect before performing any statistical analysis, which is in turn prone to errors especially if the sample groups are not split equally between batches. In this case group differences influence the batch effect reducing statistical power after the correction., the both effects are so called interdependent. The opposite could happen if the experimental design is heavily unbalanced leading to group differences induced by batch correction. Finally it is recommended to include the batch factor should in the linear model as a co factor and statistically quantify it. (Nygaard et al. 2016) However it could be of interest to observe biological relevance of the factor of interest in the dimensionality reduction visualization, so a batch correction option was included to meet this need. To address the non-biological experimental variations, also known as technical factors, the R package sva (Leek et al. 2012) with the ComBat function (Johnson et al. 2007) was used. The expression of a gene g for the batch m_i and sample j can be defined as $Y_{ijg} = a_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$, where γ and $\delta\epsilon$ represent the systematic error caused by the batch. The Bayesian framework is employed to remove this systematic error by shrinking the effect and pooling information over the features g. Ideally the effects a and β remain as natural as they are and should reflect the true biological difference between experimental obtained samples or groups. Before applying batch correction, the data is standardized using Z-transformation. By the method of moments (= the mean and the variance) the two batch effect parameters γ and $\delta\epsilon$ were estimated and can be subtracted from each value for the particular feature driving statistical moments at a comparable level between batches. The ComBat function allows the user to specify a desired effect,

which represents the actual biological variance to be preserved while removing non-biological variance. This is achieved by constructing a design matrix, which is provided as prior information about the dataset within the Bayesian framework. However, in cases of poor experimental design, the factors contributing to unwanted and desired variation may be confounded. In the worst case, if only one sample type per run is observed, it becomes challenging for the program to distinguish between batch-induced variance and biological variance. In such cases, the user is advised to adjust the experimental layout for future experiments. The detection algorithm for confounding effects is facilitated by QR matrix decomposition. The matrix $A = QR$ is decomposed into an orthogonal matrix Q and an upper triangular matrix R. By comparing the rank of matrix A with the number of biologically relevant factors, it is possible to determine the presence of confounding effects in the dataset. The rank is defined as the maximum number of linearly independent columns. The same approach is used in the limma package (Phipson et al. 2016) to identify non-solvable coefficients in experimental designs.

2.2.2.17 Dimensionality reduction After cleaning the data and crystallizing from impurities such as noise and unwanted effects, the biological differences between sample types can be obtained. Proteomic data consists of multi dimensions (=features) which are overall hard to understand by a glance. An approach to observe the differences without a closer sight of exact effect on the proteinaceous ensemble, but still keep a high level of information can be dimensionality reduction. After the cleaning procedures any additional effect caused by technical or biological replication should be removed, however it could be found be highlighting according to the particular factor. This makes the dimensionality reduction a useful tool before starting any statistical analysis by checking for involuntarily introduced bias.

2.2.2.17.1 Principle component analysis (PCA) The most common approach in data science for any kind of high dimensional data is the PCA. For the pipeline the bioconductor package scater (McCarthy et al. 2017) was used. Every principle component can be understood as a vector pointing in a specific direction derived as the eigenvector from the covariance matrix of the expression set. The covariance matrix indicates for the co linearity between the variance of all elements compared. After calculating the covariance matrix the eigenvectors for the covariance matrix will be obtained by solving the quadratic equations to obtain the eigenvalues λ. Since $Av = \lambda v$ the eigenvector v can be obtained by solving the equation for each element of the vector. The eigenvectors v never changes is direction and explains the covariance within the dataset and so

also the mathematical differences between the samples. The eigenvectors are sorted according to their eigenvalues in decreasing order and therefore decreasing explained variance. For graphical visualizations the principle components are plotted against each other. The first two components, which explain the majority of variance in the dateset highlight the difference between samples.

2.2.2.17.2 Uniform manifold approximation & projection (UMAP) UMAP (McInnes et al. 2018) is a dimensionality reduction technique that was developed in recent years and is based on Riemann geometry. It shares similarities to the t-distributed stochastic neighbor embedding (t-SNE) in terms of graphical visualization. Riemann geometry is a mathematical framework that deals with three dimensional functions using volumes, areas and vectors, which are also concept of topological data analysis. Before performing UMAP, three assumptions need to be considered:

1. There exists a manifold on which the data would be uniformly distributed. → This assumption implies that the probability for each value in the biological dataset is the same.
2. The underlying manifold of interest is locally connected. → In other words, similar samples will have a similar proteinaceous ensemble.
3. Preserving the topological structure of this manifold is the primary goal. → This means that the homeomorphic structure (the property of being able to transform one shape into another without tearing or gluing) is locally preserved, and local values in the dataset can be compared with each other.

(McInnes et al. 2018)

The UMAP algorithm begins by constructing a graph with a defined neighborhood parameter, typically denoted as k, which can be selected computationally. UMAP employs the k-nearest neighbor descent algorithm (NN-descent). Similar to the k-nearest neighbor algorithm (KNN), used in the missing value imputation part of this project, NN-descent recursively iterates to find the smallest distance to the respective neighbors within the defined neighborhood size k. Increasing the value of k gradually strengthens the clustering of the data, but at a certain point, fine structure may be lost, and only rough estimates of the underlying principle can be observed. The mathematical graph in UMAP is defined by nodes connected by edges, with each node internally connected to at least one other node. Edges are distinguished from each other not only by their connection but also by their importance, which is often referred to as weight. Whereas the weight function is calculated as $w((x_i, x_{i_j})) = \exp(\frac{-max(0,d(x_i,x_{i_j}-\rho_i)}{\sigma_i}$ and determine the parameter ρ for the weight function that:

$\rho_i = \min\left\{d(x_i, x_{i_j}) \,|\, 1 \leq j \leq k, d(x_i, x_{i_j}) > 0\right\}$. Furthermore σ will be set to meet the following conditions: $\sum_{j=1}^{k} \exp(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}) = \log_2(k)$ and acts as a normalisation factor. The obtained graph $\overline{G}$ is called the fuzzy simplicial set which is obtained by approximating the geodesic distance (= locally length-minimizing curve) of the data points, leading to a topological approximation. The second phase of the algorithm adjusts the layout of the weighted graph to a representative view, but preserves the characteristics and shows the underlying topological principle of the data. In order to obtain a visual representation of the graph, the algorithm applies repulsive forces $\frac{2b}{(\epsilon + ||y_i - y_j||_2^2)(1 + a||y_i - y_j||_2^{2b})}(1 - w((x_i, x_j)))(y_i - y_j)$ among vertices and attractive forces $\frac{-2ab||y_i - y_j||_2^{2(b-1)}}{1 + ||y_i - y_j||_2^2} w((x_i, x_j))(y_i - y_j)$ along edges. Where a and b are denoted as hyperparameters. By iterating through possible conformations until converging towards a local minimum with decreasing repulsive and attractive forces, the UMAP reaches it's aim to reveal the underlying properties of the data. The application utilizes the bioconductor package scater for performing UMAP (McCarthy et al. 2017).

2.2.2.18 Validation of the methods applied to the dataset In the context of the validation of methods applied to a dataset, the mutual information calculation is used to assess the impact of computations on the dataset and the information it carries. Mutual information, denoted as $I(X;Y)$, is a measure of the dependence or information shared between two variables, X and Y, in this case sample types. It is calculated using the equation $I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$, where H represents the entropy, $H(X|Y)$ and $H(Y|X)$ denote for the both conditional probabilities and $H(X,Y)$ for the joint entropy. Entropy, as a concept in information theory, quantifies the level of disorder or uncertainty associated with a random event. Events that are highly likely to occur are considered less informative than events that are more unexpected or less likely. In a Venn diagram $I(X;Y)$ (the mutual information) corresponds to the intersection of the two conditional entropys $H(X|Y)$ and $H(Y|X)$. To calculate the mutual information between two variables X and Y we can alternatively express the above equation as $I(X;Y) = D_{KL}(P_{(X,Y)}||P_X \otimes P_Y)$. D_{KL} denotes for the Kullback-Leibler divergence which is a concept of relative entropy between two variables and tells us how much 2 probability distributions differ from each other and $P_X \otimes P_Y$ represents the product of the marginals. The underlying principle of divergence is the integral asymmetry in Bayesian inference, which means that divergence does not satisfy the triangle inequality, and differences are not the same depending on the starting point. By comparing the mutual information before performing any calculations with the mutual information calculated after the computations, it is possible to assess whether the applied methods have significantly altered the dependence or

information content. The goal is to choose a method which preserves the information contributed by each biological sample type to the experiment. The R package "infotheo" is utilized to perform the estimates for the entropy of the two variables. Mutual information is returned as natural unit of information (=nat), a measurement proportional to the to the Shannon entropy ($1nat = \frac{1}{ln2} shannons$). Before calculating the pairwise mutual information of the expression matrix, values were discretized into bins with the discretize function of above package. The mutual information is calculated within each sample type, and it can also be calculated between a selection of sample types to determine if the computations performed affect the information content. After that all pairwise mutual information within the expression set before and after the pipeline is obtained. The pairwise differences can be observed as a boxplot for a particular sample type indicating the change in dependence within and also compared to another sample type.

2.2.2.19 Statistics

2.2.2.19.1 Differential expression After visualizing the biological tendency in the dataset, the next step is hypothesis testing. Several experimental designs can be tested against the null hypothesis h_0, which states that there is no significant difference between the sample types. Before proceeding with any statistical testing, it is important to observe the data using either a quantile-quantile norm plot (qq-norm plot) or a histogram. The quantile-quantile plot (qq-plot) serves a vital role in exploratory statistics by visualizing whether two datasets follow the same underlying distributions. In the case of statistical preparation for parametric tests, the first distribution is the dataset of interest, while the second dataset is the normal distribution. If the dataset follows an exact normal distribution, the values will align with the red line in the chart. However, values at the ends of the distribution may deviate from the normality line, indicating differentially expressed genes. These genes can be observed both in the histogram and the qq-norm plot.

Parametric statistical tests rely on the assumption of normality in the data, which needs to be achieved before building a model. To achieve normality, it is recommended to try various combinations of transformation and normalization methods. By observing the results in the plots, one can determine which combination of data processing methods satisfies the conditions required for the tests.

2.2.2.19.2 Differential expression To test for differential expression, the analysis utilizes the "limma" package from R Bioconductor (Phipson et al. 2016). This package employs a linear model

approach to determine the fold expression between sample groups. Before initiating the analysis, two matrices need to be computed. The design matrix identifies samples based on their sample type and defines the experimental design. An algorithm within the program's backend logic is utilized to automatically generate this matrix according to the user's selection in the interface. The expression matrix contains intensities for each identified protein and is obtained at the end of the pipeline. By calling a function with the design and expression matrices as arguments, the contrast matrix is calculated, enabling the user to decide which comparisons to consider and also perform tests on multiple factors. When observing multiple batches at once, as already explained in the chapter batch correction this effect needs to be taken into consideration when performing differential expression analysis. This is done by the multi factor option in the user interface by selecting the batch as a co factor and inclusion in the design matrix. The advantage of this procedure is, that we can quantify the biological relevance of working in multiple batches and avoid publishing false positive results. The multi factor option uses an additive model as a design matrix, whereas continuous variables (such as size, weight etc.) can also be used to build the statistical model.

The subsequent step involves creating a linear model between groups using log-ratios of their expression values. In a two-way design, the expression of the gene y can be explained as $Exp_Y = \beta_0 + \beta_1 X_1 + \epsilon$, where β_0 represents the intercept, which can be interpreted as the mean expression of the gene. β_1 represents difference in mean of the treatment (or condition) on the discrete or continuous variable X_1, and ϵ acts as an error term. The interface allows for additive models with multiple factors, allowing the user to add factors according to various scenarios. For example, in a two-factor model, it would look like: $Exp_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. The principle remains the same for models with multiple factors. When visualizing the linear model, the x-axis would represent the gene for the sample types or other factors, whereas the y-axis would account for the expression value. By drawing a line through the cloud of data points the model is constructed by minimizing the quadratic distance to every data-point, called the least-square method or also ordinary least-square method. With the regression line the above mentioned equation can be constructed.

Next the statistical module of the program performs an empirical Bayes model utilizing the eBayes (Smyth 2004) function of the Limma package. The eBayes function utilizes a moderated t-statistics following t-distributed values. The advantage of this method over the posterior odds is the reduction of hyperparameters to estimate used for the modelling process, also called shrinkage. In this context, the hyperparameters to estimate are the coefficient β_{gj} for gene g and sample i, as well as the variance σ_g^2 across genes g. In the Bayesian approach, probabilities are updated after obtaining new data, referred to as conditional probabilities. For estimation, it is assumed that $\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$,

representing the prior information of the model. The probability that $\beta_{gj} \neq 0$ is denoted as p_j, where p_j represents the expected proportion of truly differentially expressed genes. The expected distribution of log-fold changes follows the distribution $\beta_{gj}|\sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$. The posterior mean of σ_g^2 given s_g^2 is calculated as $s_g^{-2} = E(\sigma_g^2|s_g^2) = \frac{d0s_0^2+d_g s_g^2}{d_0+d_g}$. This is where the shrinkage occurs, as the posteriors affect the prior values based on their respective sizes and degrees of freedom. The moderated t-statistic is defined as $\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\hat{s}_g\sqrt{v_{gj}}}$. Once the values of $\tilde{t}$ and s^2 are calculated, the posterior odds are computed, providing the odds that a particular gene is differentially expressed. The odds for gene g being differentially expressed are denoted as $O_{gj} = \frac{p(\beta_{gj}\neq 0|\tilde{t}_{gj}, s^2 g)}{p(\beta_{gj}=0|\tilde{t}_{gj}, s^2 g)}$, where the numerator represents the probability that the gene is differentially expressed, and the denominator represents the probability that the gene is not differentially expressed (Smyth 2004).

Limma outputs a table with all proteins, their respective logfold change, p-values and adjusted p-values. The user can choose different methods for the correction of the p-values, which is necessary when testing the same hypothesis multiple times. A common method in omics studies is the false discovery rate (=fdr) by Benjamini Hochberg (Benjamini & Hochberg 1995). By ranking all the p-values from smallest to largest the p-values are adjusted sequentially using the formula $p_(i) = \frac{r_i}{m}Q$. The rank in the p-value table is denoted as r_i , m reflects the total number of tests and Q represents false discovery rate which is set to 5%. The second option to choose is the Benjamini & Yekutieli (=BY) (Benjamini & Yekutieli 2001)method, which is a further development of the fdr by the same statistician and has a wider range of application based on the dependencies for the test. The third and most conservative option is the Holm method (Holm 1979) which controls the error sequentally in a family wise manner, which is considered as traditional in regards to the above mentioned experimental wise error rates. The above mentioned correction methods are ordered in the way from the most liberal to the most conservative one. The selection is supposed to give the user the possibility questioning results on multiple levels.

2.2.2.20 Protein set enrichment analysis After differential expression the upcoming question could be, in which context the up or down regulated proteins are. Enriched proteins can be understood as an overabundance within a specific set. This set can be a pathway or a selection of common denominators in biological contexts, such as cell types or oncological factors. The statistical testing to identify the overabundant proteins of a specific set is similar to a χ^2 test developed by Karl Pearson (Pearson 1900) or the exact Fisher test (Sprent 2011) for tables with 4 fields in small sample sizes. Within the list are all truly differentially expressed genes, an annotation is the identification of a protein in the desired biological context. We expect that the values populate the fields in the table

as follows:

	in list	not in list	totals
with annotation	(A+B)(A+C)/N	(A+B)(B+D)/N	A+B
without annotation	(C+D)(A+C)/N	(C+D)(B+D)/N	C+D
	A+C	B+D	N

Under the null hypothesis H_0 states that the distribution and population of the table is random and there is no clear tendency with significant evidence. Against that the alternative hypothesis H_A says that there is an underlying tendency the table is populated. This leads to our test statistics, the observed values (=O) in the table are significantly different to the expected (=$E = \frac{(A+B \vee C+D)}{N}$ ones. One approach is approximating $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ and deriving the p-value by the area under the density curve of the χ^2 distribution $Q = \sum_{i=1}^{k} Z_i^2$. In the distribution k is defined by the degrees of freedom (=df), what can be obtained by $k = (columns - 1) * (rows - 1) = df$ of our table. Another approach to obtain an exact p-value would be Fisher's exact test, which states that the margins are distributed according to the hypergeometric distributions. The test statistic leads directly to the p-value with no need for approximation: $p = 1 - \sum_{i=0}^{A-1} \frac{\binom{B}{i}\binom{D-B}{A+C-i}}{\binom{D}{A+C}}$ Here D denotes either for the background distribution which can be the total number of proteins obtained in our experiment, all proteins with annotation or a custom selected background. These three options are implemented in the user interface. After obtaining the p-values for each protein set, the subsequent critical step involves the correction for multiple testing, considering the inherent risk of false positive results. Multiple testing correction methods adjust the significance thresholds to account for the increased probability of detecting false positives when performing multiple statistical tests. In this protein set enrichment analysis, the user is provided with a selection of established correction methods, akin to those commonly utilized in the field of differential expression analysis, ensuring robust and reliable interpretation of the results.

2.2.2.20.1 Pathway based protein enrichment analysis The pathway-based protein enrichment analysis is performed using the R Bioconductor package clusterProfiler (Wu et al. 2021). The analysis utilizes the "Kyoto Encyclopedia of Genes and Genomes" (KEGG) database, which is accessed through a function that maps the UniProt IDs to corresponding pathways. Subsequently, the Fisher's exact test is applied to identify significantly enriched pathways. To focus on truly differentially expressed proteins, users have the flexibility to specify p-value and fold change cutoffs. Furthermore, the method for p-value correction can be selected based on user preference. Once the

enriched protein pathways are determined, they can be visually represented using bar or dotplots, providing a clear and intuitive representation of the data.

2.2.2.20.2 Ontology based protein enrichment analysis

3 Results

Some more guidlines from the School of Geosciences.

This section should summarise the findings of the research referring to all figures, tables and statistical results (some of which may be placed in appendices). - include the primary results, ordered logically - it is often useful to follow the same order as presented in the methods. - alternatively, you may find that ordering the results from the most important to the least important works better for your project. - data should only be presented in the main text once, either in tables or figures; if presented in figures, data can be tabulated in appendices and referred to at the appropriate point in the main text.

Often, it is recommended that you write the results section first, so that you can write the methods that are appropriate to describe the results presented. Then you can write the discussion next, then the introduction which includes the relevant literature for the scientific story that you are telling and finally the conclusions and abstract – this approach is called writing backwards.

4 Discussion

the purpose of the discussion is to summarise your major findings and place them in the context of the current state of knowledge in the literature. When you discuss your own work and that of others, back up your statements with evidence and citations. - The first part of the discussion should contain a summary of your major findings (usually 2 – 4 points) and a brief summary of the implications of your findings. Ideally, it should make reference to whether you found support for your hypotheses or answered your questions that were placed at the end of the introduction. - The following paragraphs will then usually describe each of these findings in greater detail, making reference to previous studies. - Often the discussion will include one or a few paragraphs describing the limitations of your study and the potential for future research. - Subheadings within the discussion can be useful for orienting the reader to the major themes that are addressed.

5 Conclusion

The conclusion section should specify the key findings of your study, explain their wider significance in the context of the research field and explain how you have filled the knowledge gap that you have identified in the introduction. This is your chance to present to your reader the major take-home messages of your dissertation research. It should be similar in content to the last sentence of your summary abstract. It should not be a repetition of the first paragraph of the discussion. They can be distinguished in their connection to broader issues. The first paragraph of the discussion will tend to focus on the direct scientific implications of your work (i.e. basic science, fundamental knowledge) while the conclusion will tend to focus more on the implications of the results for society, conservation, etc.

6 Bibliography

Aebersold, R. & Mann, M. (2003), 'Mass spectrometry-based proteomics', *Nature* **422**, 198–207.

Bacharach, M. (1965), 'Estimating nonnegative matrices from marginal data', *International Economic Review* **6**, 294.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *The Annals of Statistics* **29**.

Budnik, B., Levy, E., Harmange, G. & Slavov, N. (2018), 'Scope-ms: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation', *Genome Biology* **19**, 161.

Cox, J. & Mann, M. (2008), 'Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nature Biotechnology* **26**, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. (2011), 'Andromeda: A peptide search engine integrated into the maxquant environment', *Journal of Proteome Research* **10**, 1794–1805.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**, 65–70.

Houtven, J. V., Hooyberghs, J., Laukens, K. & Valkenborg, D. (2021), 'Constand: An efficient normalization method for relative quantification in small- and large-scale omics experiments in r bioconductor and python', *Journal of Proteome Research* **20**, 2151–2156.

Johnson, W. E., Li, C. & Rabinovic, A. (2007), 'Adjusting batch effects in microarray expression data using empirical bayes methods', *Biostatistics* **8**, 118–127.

Lan, L., Djuric, N., Guo, Y. & Vucetic, S. (2013), 'Ms-k nn: protein function prediction by integrating multiple data sources', *BMC Bioinformatics* **14**, S8.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. (2012), 'The <tt>sva</tt> package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics* **28**, 882–883.

Maes, E., Hadiwikarta, W. W., Mertens, I., Baggerman, G., Hooyberghs, J. & Valkenborg, D. (2016), 'Constand : A normalization method for isobaric labeled spectra by constrained optimization', *Molecular & Cellular Proteomics* **15**, 2779–2790.

Marx, V. (2019), 'A dream of single-cell proteomics', *Nature Methods* **16**, 809–812.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. (2017), 'Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r', *Bioinformatics* **33**, 1179–1186.

McInnes, L., Healy, J. & Melville, J. (2018), 'Umap: Uniform manifold approximation and projection for dimension reduction'.

Nygaard, V., Rødland, E. A. & Hovig, E. (2016), 'Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses', *Biostatistics* **17**, 29–39.

Pearson, K. (1900), 'X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**, 157–175.

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. (2016), 'Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression', *The Annals of Applied Statistics* **10**.

Sakia, R. M. (1992), 'The box-cox transformation technique: A review', *The Statistician* **41**, 169.

Senko, M. W., Beu, S. C. & McLaffertycor, F. W. (1995), 'Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions', *Journal of the American Society for Mass Spectrometry* **6**, 229–233.

Smyth, G. K. (2004), 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical Applications in Genetics and Molecular Biology* **3**, 1–25.

Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., Kharchenko, P., Koller, A. & Slavov, N. (2021), 'Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using scope2', *Genome Biology* **22**, 50.

Sprent, P. (2011), *Fisher Exact Test*, Springer Berlin Heidelberg.

Tannous, A., Boonen, M., Zheng, H., Zhao, C., Germain, C. J., Moore, D. F., Sleat, D. E., Jadot, M. & Lobel, P. (2020), 'Comparative analysis of quantitative mass spectrometric methods for subcellular proteomics', *Journal of Proteome Research* **19**, 1718–1730.

Vanderaa, C. & Gatto, L. (2021), 'Replication of single-cell proteomics data reveals important computational challenges', *Expert Review of Proteomics* **18**, 835–843.

Wilm, M. (2011), 'Principles of electrospray ionization', *Molecular & Cellular Proteomics* **10**, M111.009407.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X. & Yu, G. (2021), 'clusterprofiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation* **2**, 100141.

7 Appendix(ces)

7.1 Appendix A: additional tables

Insert content for additional tables here.

7.2 Appendix B: additional figures

Insert content for additional figures here.

7.3 Appendix C: code

Insert code (if any) used during your dissertation work here.