
# 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.

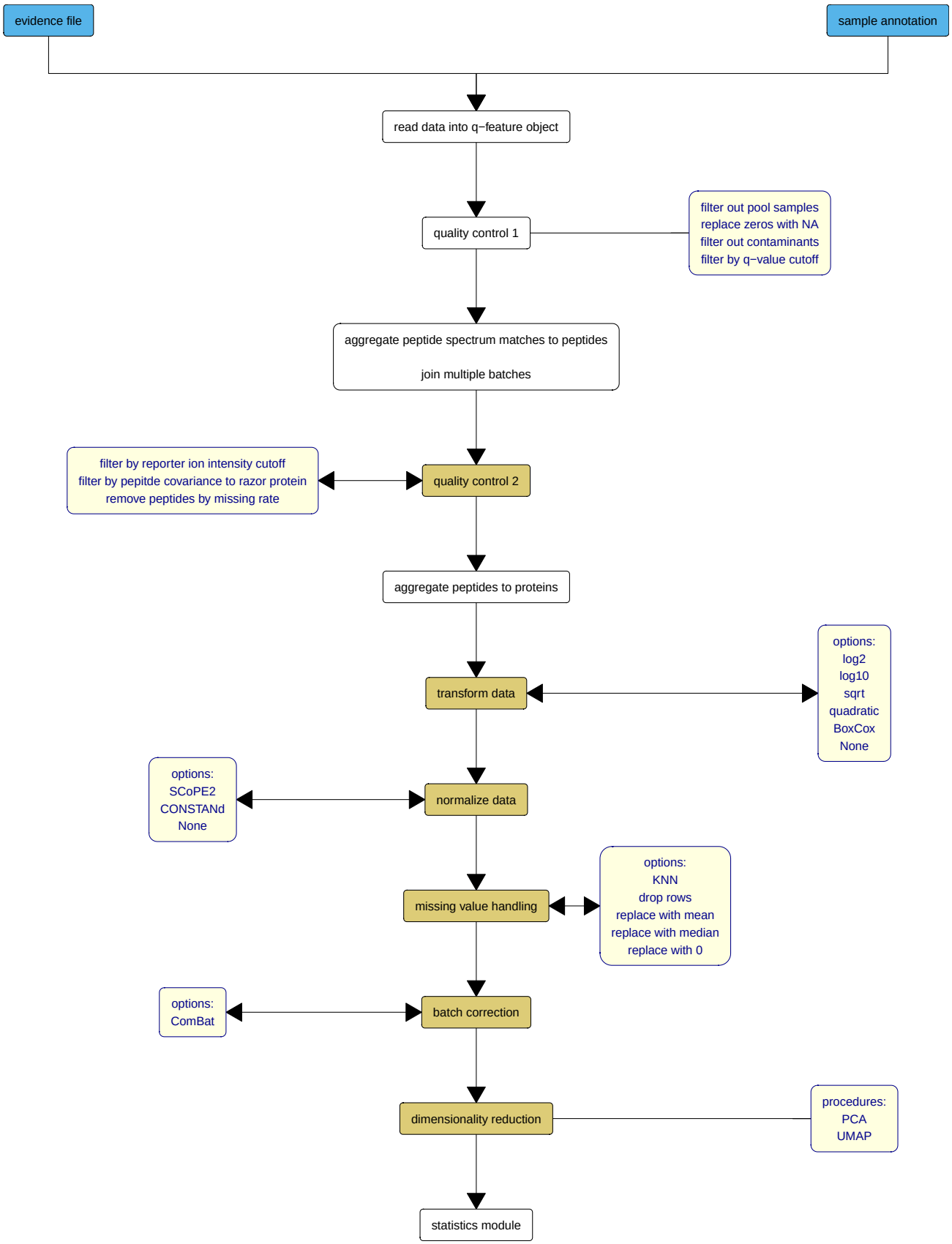
**2.2.1.5 Improving peptide mass accuracy** The intensity-weighted average of the ms peak centroids (as described in the 3D peak identification) refers to the mass of the peptide. Corrections highly depend on the analyzer, MaxQuant uses for an orbitrap typed analyzer a correction value of 1ppm. Autocorrelation 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
filter by pepitde covariance to razor protein
remove peptides by missing rate
quality control 2
aggregate peptides to proteins
transform data
options:
log2
log10
sqrt
quadratic
BoxCox
None
options:
SCoPE2
CONSTANd
None
normalize data
missing value handling
options:
KNN
drop rows
replace with mean
replace with median
replace with 0
options:
ComBat
batch correction
dimensionality reduction
procedures:
PCA
UMAP
statistics module

**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 recursion on the data to determine the statistical parameter lambda.

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.

Further filter steps can include correction between multiple runs. This kind of process need the addition of a reference channel. However when observing a single run, these steps are not crucial for the upcoming analysis.

#### 2.2.2.14 Downstream Analysis

##### 2.2.2.14.1 Principle component analysis (PCA)

Protein levels can be projected to their principle components (PC) and clustered to their specific cell type. So cell types are distinguished by

their proteinaceous composition.

**2.2.2.15 Testing for differential expression** In order to test for differential expression the package limma from R bioconductor was used. (Phipson et al. 2016). The package uses a linear model approach to define a fold expression between sample groups. Before starting the analysis two matrices need to be computed. The design matrix identifies samples according to the sample type and defines the experimental design. In order to create the matrix automatically a simple algorithm is used within the programmed backend logic. The expression matrix contains intensities for each identified protein will be obtained at the end of the pipeline. A function call with the arguments design and expression matrix calculates the contrast matrix. The next step is creating a linear model between groups by log-ratios of their expression values.

# 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.

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.

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

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.

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.

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.

# 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.