# EXPERIMENTAL DESIGN AND DATA-ANALYSIS IN LABEL-FREE QUANTITATIVE LC/MS PROTEOMICS: A TUTORIAL WITH MSQROB

This article also constitutes section 9.2. of my PhD thesis.

**When referring to this work, please always cite the original source at:**

Goeminne L.J.E., Gevaert K. and Clement L. (2018). **Experimental design and data-analysis in label-free quantitative LC/MS proteomics: A tutorial with MSqRob.** *Journal of Proteomics*. 171(Supplement C), 23-36

https://www.sciencedirect.com/science/article/pii/S1874391917301239

## Personal summary

This article is published as an invited tutorial paper in which we outline key statistical concepts to help researchers to design proteomics experiments and showcases of quantitative proteomics data analysis with MSqRob, our R software package for improved differential protein abundance analysis in label-free MS-based proteomics. MSqRob is freely available on GitHub (https://github.com/statOmics/MSqRob) and is implemented in a "Shiny" user-friendly graphical interface. For this manuscript, I designed and performed analyses, set up the GitHub repository and wrote the paper together with my supervisors.

## Highlights

- Complex experiments and lack of convenient software makes MS-based label-free proteomics data analysis challenging
- We provide key experimental design concepts and data analysis guidelines
- The MSqRob package combines legitimate statistical modeling for relative protein quantification from peptide-level data with an easy-to-use graphical interface
- We show hands-on with two worked examples how to use the MSqRob graphical user interface
- Scripts to run MSqRob in bash mode are provided at https://github.com/statOmics/MSqRob

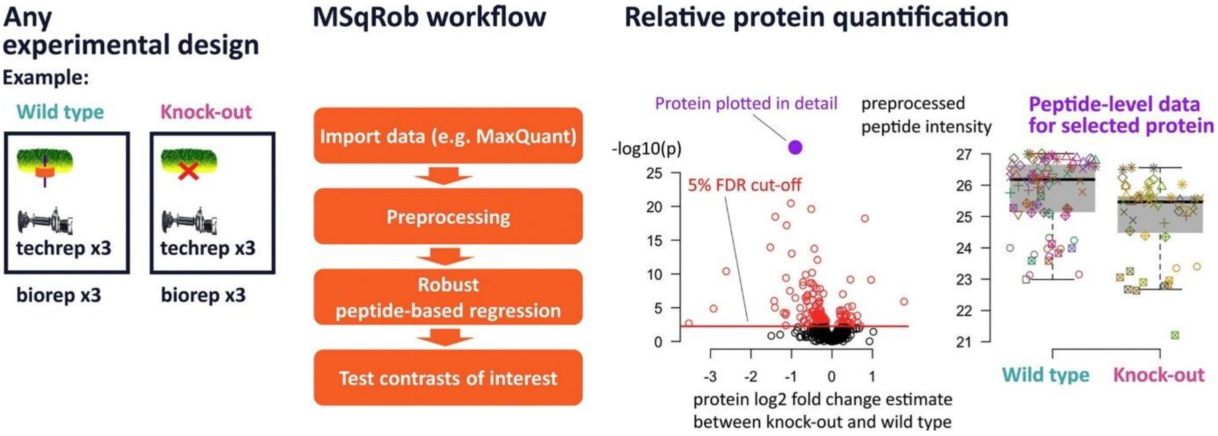
## Abstract

Label-free shotgun proteomics is routinely used to assess proteomes. However, extracting relevant information from the massive amounts of generated data remains difficult. This tutorial provides a strong foundation on analysis of quantitative proteomics data. We provide key statistical concepts that help researchers to design proteomics experiments and we showcase how to analyze quantitative proteomics data using our recent free and open-source R package MSqRob, which was developed to implement the peptide-level robust ridge regression method for relative protein quantification described by Goeminne et al. MSqRob can handle virtually any experimental proteomics design and outputs proteins ordered by statistical significance. Moreover, its graphical user interface and interactive diagnostic plots provide easy inspection and also detection of anomalies in the data and flaws in the data analysis, allowing deeper assessment of the validity of results and a critical review of the experimental design. Our tutorial discusses interactive preprocessing, data analysis and visualization of label-free MS-based quantitative proteomics experiments with simple and more complex designs. We provide well-documented scripts to run analyses in bash mode on GitHub, enabling the integration of MSqRob in automated pipelines on cluster environments (https://github.com/statOmics/MSqRob).

## Significance

The concepts outlined in this tutorial aid in designing better experiments and analyzing the resulting data more appropriately. The two case studies using the MSqRob graphical user interface will contribute to a wider adaptation of advanced peptide-based models, resulting in higher quality data analysis workflows and more reproducible results in the proteomics community. We also provide well-documented scripts for experienced users that aim at automating MSqRob on cluster environments.

## Graphical abstract


Any experimental design
Example:
Wild type
Knock-out
techrep x3
techrep x3
biorep x3
biorep x3
MSqRob workflow
Import data (e.g. MaxQuant)
Preprocessing
Robust peptide-based regression
Test contrasts of interest
Relative protein quantification
Protein plotted in detail
-log10(p)
5% FDR cut-off
25
20
15
10
5
0
-3
-2
-1
0
1
protein log2 fold change estimate between knock-out and wild type
preprocessed peptide intensity
27
26
25
24
23
22
21
Peptide-level data for selected protein
Wild type
Knock-out

## Keywords

Differential protein abundance; Biostatistics; Label-free quantification; Tandem mass spectrometry; Experimental design; Peptide-based linear model

## Historical background

Proteomics was revolutionized with the rise of biological mass spectrometry, genome sequencing and bioinformatics [1]. In a typical workflow for comprehensive proteome analysis, proteins are digested with specific proteases such as trypsin. The resulting peptide mixtures are then separated by high performance liquid chromatography (HPLC). Subsequently, a mass spectrometer coupled to the HPLC instrument is used to analyze the eluting peptides and the generated (tandem) mass spectra are mapped to theoretical spectra generated based on protein sequences stored in databases.

In early days, MS-based proteomics was used to just identify proteins. As technology matured, quantitative information was extracted from proteome samples. Efforts have been made to determine absolute protein amounts based on mass spectra. These can be very sensitive in a targeted proteomics context [2], [3], [4] but current methods for proteome-wide absolute quantification remain rather crude due to massive ionization efficiency differences between peptides [5].

In this tutorial the focus is on relative quantification; i.e. the abundance of a given protein is compared over different samples. One of the first relative quantification technologies was based on isotope-coded affinity tags (ICAT) [6]. Later, metabolic labeling with stable isotopes, e.g. $^{15}$N and SILAC, emerged, where some samples were grown in medium made from the most abundant natural isotopes, and other samples in medium containing stable heavy isotopes [7], [8], [9]. Note that metabolic labeling can be rather expensive and is mainly performed on in vitro cell cultures. In cases where metabolic labeling is not possible, post-metabolic isobaric multiplex labeling such as iTRAQ [10] and tandem mass tags (TMT) [11] can be used. However, post- or non-metabolic labeling may be incomplete, leading to higher sample-to-sample variability compared to metabolic labeling. More information on quantification with isobaric labeling can be found in Rauniyar and Yates [12]. Labeling has the intrinsic advantages that both the analytical time as well as the run-to-run variation are reduced as because it enables sample multiplexing in one MS-run as two or more peaks can be measured in the same MS- or MS/MS-spectrum.

Nowadays, label-free methods are becoming more and more standard. Such methods scale very well, have no real upper limit on the number of samples that can be compared (even in retrospect) and bypass the labor-intensive and often expensive sample labeling steps. Moreover, up to 60% more proteins can be identified, and this at a higher dynamic range because the mass spectrometer does not have to fragment each labeled form of the same peptide [13], [14]. A disadvantage of label-free quantification is that a peptide selected for fragmentation in one run might not be selected for fragmentation or result in a poorer quality MS$^2$ spectrum in another run, leading to missing values. “Match between runs” algorithms, where unidentified MS$^1$ peaks are matched to identified peaks using a tight retention time and mass/charge window, significantly improve the number of identified peaks [15]. In the remainder of the manuscript we will focus on data-dependent label-free MS-based quantification.

In data-dependent acquisition (DDA), a software identifies multiply charged peptide precursor ions with the highest intensities from deconvoluted MS (or MS$^1$) spectra. In a next step, such peptide ions are individually selected and further fragmented, typically by collision-induced dissociation, whereby MS/MS (or MS$^2$) spectra are recorded. The frequency by which peptide ions are fragmented depends on both the LC resolution and the scanning speed of the mass spectrometer, with increasingly faster instruments now mapping larger fractions of the expressed proteome than ever before [16], [17]. Fig. 1 gives an overview of a contemporary label-free mass spectrometry-based proteomics workflow.

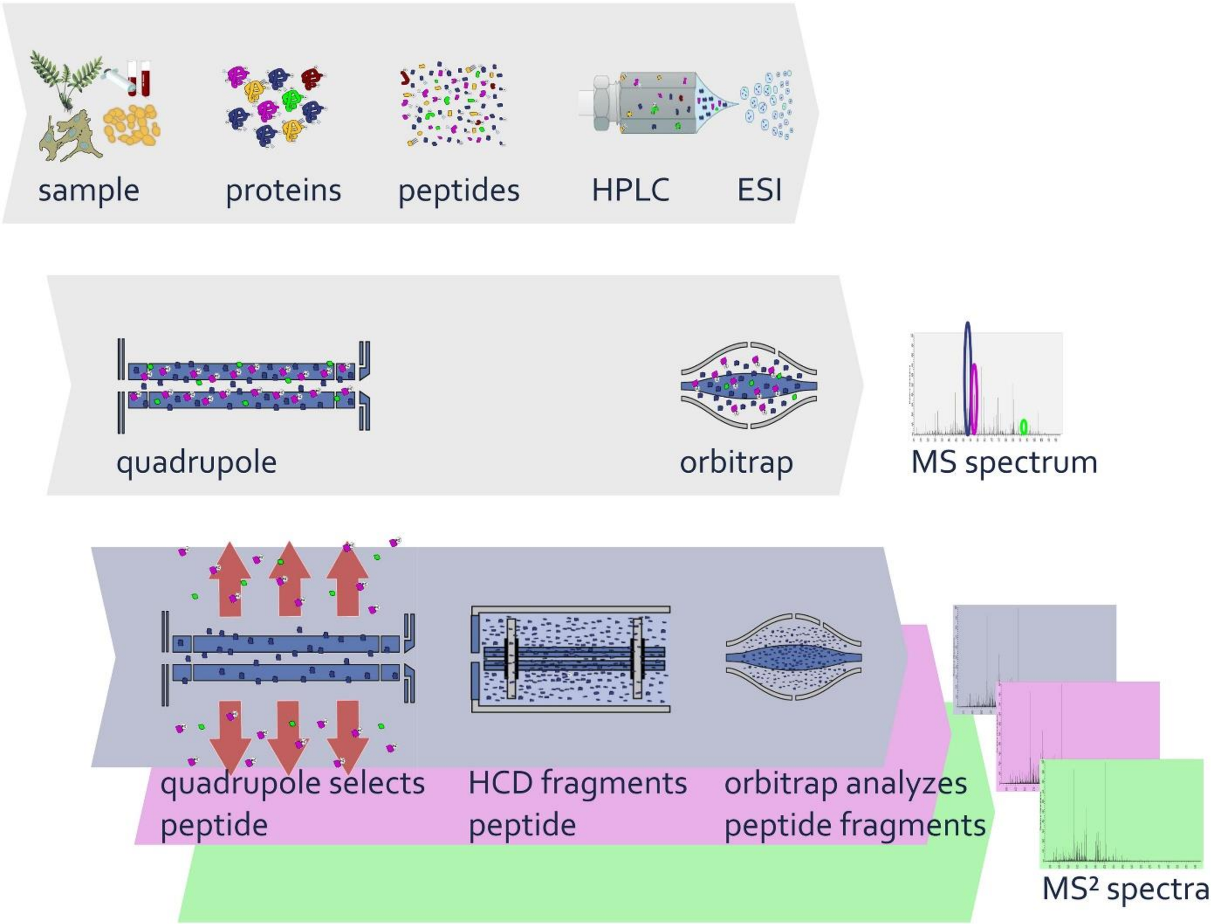


**Figure 1.** A typical DDA shotgun proteomics workflow using a quadrupole Orbitrap instrument. Extracted proteins are enzymatically digested to peptides, using a specific protease such as trypsin [18]. Peptides are then separated over a reverse-phase column and eluting peptides are transformed to gas-phase ions by electrospray ionization (ESI) [19]. At discrete time points, the eluting set of ionized peptides are

sent through the mass spectrometer and an MS spectrum is taken. Peak intensities in the MS spectrum are a proxy for peptide abundance. Upon deconvoluting the spectrum, the software identifies the highest peaks. For the next set of ionized peptides, only one peptide family present at the mass-to-charge ratio corresponding to one of the highest peaks in the MS spectrum will be separated from the rest in the quadrupole. This peptide is further fragmented in a collision-induced dissociation (CID) [20] or higher-energy collisional dissociation (HCD) [21] cell and an MS$^2$ spectrum of its fragments is taken. MS$^2$ spectra for other peptides with high MS spectral intensities are also recorded. After recording a pre-specified number of MS$^2$ spectra, the eluate composition will have changed, and a new MS spectrum is taken, followed by new MS$^2$ spectra, and so on.

Protein identification is a first important step in the data-analysis workflow. As technology advanced, manual inspection of the sheer number of MS$^2$ spectra became practically impossible (the newest generation of machines identify around 25,000 peptides from their MS$^2$ spectra in a given run, but do note that this number depends on machine settings and sample complexity). Bioinformatics software was introduced that is capable of identifying a peptide from its fragmentation spectrum given a database in which protein sequences are stored. The first one was PeptideProphet [22], using the SEQUEST algorithm [23]. The Mascot search engine introduced probability-based scoring, giving researchers a way to remove unreliable identifications at a predefined false discovery rate level [24]. Other search algorithms can be easily executed using SearchGUI [25], a graphical user interface that allows for searches with X!Tandem [26], MS-GF + [27], MS Amanda [28], MyriMatch [29], Comet [30], Tide (a fast implementation of the SEQUEST algorithm) [31], Andromeda [32], OMSSA [33], Novor [34] and DirecTag [35]. Further, tools such as PeptideShaker [36] can be used to combine results of different search engines to boost identifications. The MaxQuant search engine, which uses the Andromeda algorithm, is very popular nowadays thanks to its user-friendly graphical user interface [37]. As soon as it became possible to automatically search spectra for peptides in a database, the need for data storage, processing and visualization software also emerged [38].

Upon identification, a subsequent protein quantification step is required, which remains a tedious task for several reasons. First, sample preparation needs to be tightly controlled in order to reduce variability in protein extraction and digestion [39]. Second, the actual protein sequence surrounding the protease recognition site as well as protein modifications influence a protease's cleavage efficiency, thus possibly yielding peptides at varying levels [40]. Third, some peptides from a given protein ionize poorly, while others give very strong signals, depending on the peptide sequence and its modification status [41]. Fourth, some peptides, so-called razor peptides, cannot be uniquely attributed to a single protein and should thus either be used with extreme care when quantifying a protein or excluded altogether. Fifth, mass spectrometers are stochastic, thus sampling of MS$^1$ spectra is inherently discrete, whereas peptides continuously elute from the column; hence, the observed peptide peak intensities may vary between samples. Sixth, competition for ionization between co-eluting peptides causes extra variability [42], [43], [44]. And, finally, co-eluting peptides with similar mass-to-charge ratios may be co-fragmented, resulting in chimeric spectra and biased quantifications [45], [46], [47].

Relative quantification can be done either through **spectral counting** or through **intensity-based methods** (see Blein-Nicolas and Zivy [48] for a complete overview), although a few methods, like ProPCA [49], combine both approaches. Spectral counting consists of comparing the number of peptide-to-spectrum matches (PSMs; this includes all redundant peptide identifications due to modifications, charge states and expiration of dynamic exclusion) for a protein across samples as a proxy for protein abundance [50]. While this technique has the advantage of its simplicity and is able to quantify proteins for which no peptides are found in one condition, it has become rather obsolete for MS-based quantification as precision can be an issue, especially when comparing small differences in abundance [51]. Also, dynamic exclusion settings of the mass spectrometer (i.e. the same MS peak is fragmented only once in order to boost the number of identifications) might obscure the relationship between the number of counts and protein abundance [51], [52], [53]. Further, when machine settings are changed, runs become incomparable.

Intensity-based methods make use of the more accurate information present in spectral intensities or areas under the peaks in either MS or MS/MS spectra, which causes intensity-based methods to be more sensitive [54]. Such methods can be subdivided into MS$^2$ and MS$^1$ intensity-based methods. MS$^2$ methods are less accurate as peptide fragmentation does not always occur at the maximum of the elution peak [55]. Within MS$^1$ intensity-based methods, there are broadly two approaches, which we refer to as summarization-based methods and peptide-based models.

**Summarization-based methods** comprise all methods that summarize observed peptide intensities at the protein-level before performing a statistical analysis on protein abundance [56], often in an ad hoc manner [57]. Examples include, but are not limited to summing up peptide intensities [58], [59], (weighted or trimmed) mean summarization [60], [61], (weighted or trimmed) median summarization [62] and summarization based on peptide ratios (e.g. the method developed by Dost et al. [63] and maxLFQ [64]). All but the most efficient of these (such as the ratio-based approaches and ProPCA) ignore the fact that peptide ionization efficiency strongly influences the finally reported protein intensity, which leads to a bias due to different peptides that are missing in different samples. Also, none of these methods account for the fact that for the same protein, a different number of peptides might be identified in each sample, leading to differences in precision of the summarized protein expression value. The strong correlation between a peptide's intensity and its identification probability further exacerbates these issues.

**Peptide-based models** estimate protein fold changes (FC) directly from peptide intensities within the framework of a statistical (linear) regression model. Examples include linear mixed effect models such as presented in Daly et al. [65] and Clough et al. [57] (implemented in the MSStats package [66]), but also non-linear models [67], models handling peptides that are shared between protein groups, such as the method developed by Blein-Nicolas et al. [68] and SCAMPI [69] (implemented in the protiq R package) as well as censored regression models for missing peptides such as SALPS [70] and the method developed by Karpievitch et al. [71] (implemented in the DanteR R package [72]). We and others have shown that peptide-based models outperform summarization-based methods by reducing bias and increasing sensitivity, specificity, accuracy and precision [57], [73]. However, traditional peptide-based models still suffer from (1) overfitting, (2) unstable variances and (3) outliers. Our proteomics quantification package MSqRob tackles these issues by building upon (1) ridge regression, (2) borrowing information across proteins and (3) down-weighing outliers, all of which were discussed in Goeminne et al. [74]. In this tutorial paper, we focus on the integration of peptide-based models from the MSqRob framework in current quantitative proteomics workflows.

## Basic concepts

The actual design of an experiment strongly impacts the data analysis and its power to discover differentially abundant proteins. Therefore, we first cover some basic concepts on experimental design. Next, we provide a general step-by-step overview of a typical quantitative proteomics data analysis workflow.

## Basic concepts on experimental design

The monthly column “Points of significance” in *Nature Methods* is a useful primer on statistical design for researchers in life sciences to which we extensively refer in this section (http://www.nature.com/collections/qghhqm/pointsofsignificance).

For proteomics experiments it is important to differentiate between **experimental units** and **observational units**. Experimental units are the subjects/objects on which one applies a given treatment, often also denoted as biological repeats. In a proteomics experiment, the number of experimental units is typically rather limited (e.g. three biological repeats of a knockout and a wild-type sample). The measurements, however, are applied on the observational units. In a shotgun proteomics experiment, these are the individual peptide intensities. For many proteins, there are thus multiple observations/peptide intensities for each experimental unit, which can be considered as technical replicates or pseudo-replicates [75]. Hence, one can make very precise estimates on the technical variability of the intensity measurements; i.e. how strongly intensity measurements fluctuate for a particular protein in a particular sample. However, the power to generalize the effects observed in the sample to the whole population remains limited as most biological experiments typically only have a limited number of biological repeats [76]. We thus strongly advise researchers to think upfront about their experimental design and to maximize the number of biological repeats as much as feasible (we suggest at least three, and preferably more).

Another important concept is that of **blocking** [77], which randomizes the different treatments to experimental units that are arranged within groups/blocks (e.g. batches, time periods) that are similar to each other. Due to practical constraints, it is often impossible to perform all experiments on the same day, or even on the same HPLC column or mass spectrometer, leading to unwanted sources of technical variation. In other experiments, researchers might test the treatment in multiple cultures or in big experiments that involve multiple labs. A good experimental design aims to mitigate unwanted sources of variability by including all or as many treatments as possible within each block. That way, variability between blocks can be factored out from the analysis when assessing treatment effects (Fig. 2).

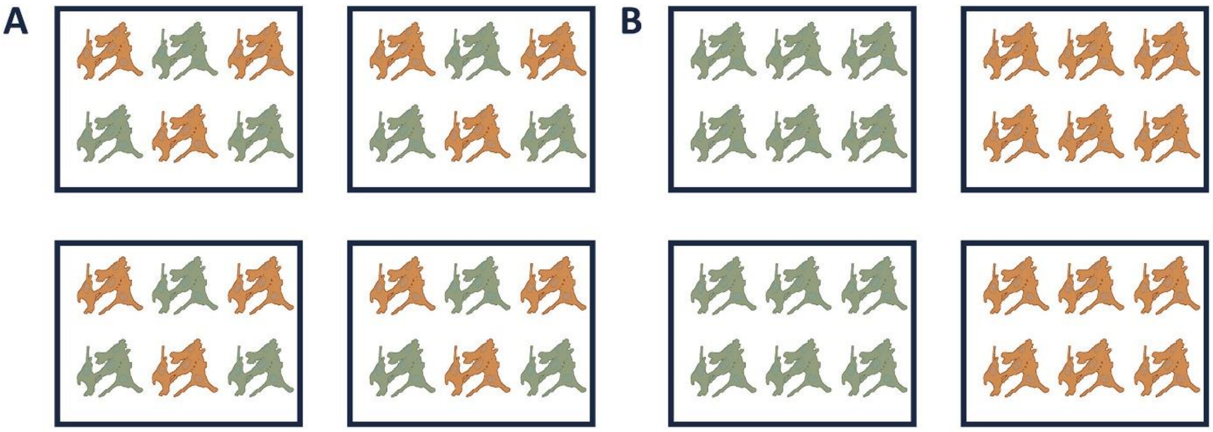


**Figure 2.** Example of a good (A) and a bad (B) design. In design A, both the green and orange treatments are divided equally within each block. That way, the treatment effect can be estimated within a block. In design B, each block contains only one treatment, so the treatment effect is entirely confounded with the blocking effect and it is thus impossible to draw meaningful conclusions on the treatment (unless one would be willing to assume that the blocking effect is negligible, which is a very strong assumption that cannot be verified based on the design).

Finally, it is important to correctly account for the degree of pseudo-replication within each block so as to provide final FC estimates with correct standard errors. Note that pseudo-replication always occurs in peptide-based linear models because the peptides from the same protein in the same sample can be considered as technical replicates for protein expression.

### *A general label-free quantitative proteomics data analysis workflow*

A first crucial step is the identification of peptides from mass spectra. Peptide identification has already been touched upon in the historical background section. Methods for peptide identification are constantly being developed and improved. Providing a complete review on the strengths and limitations of popular search engines is outside the scope of this tutorial.

Once peptides are identified and their spectral intensities are determined, the identified peptide-to-spectrum matches need to be assigned to the correct protein to perform quantification at the protein level. Often, this is trivial, but some peptides can originate from multiple proteins (so-called **razor peptides**). How to handle razor peptides is a matter of debate [78], but as their intensities might represent a combined intensity, it might be safer to remove them from the data altogether. When two or more proteins are very similar in their amino acid sequence, it can be more convenient to group them together into a "**protein group**". MaxQuant does this automatically [15]. For the remainder of this text, the term "protein" encompasses both "proteins" and "protein groups", unless explicitly stated otherwise.

Before proteins can be quantified, the intensities of identified peptides need to be preprocessed. Raw (summarized) peptide intensities found in MaxQuant's peptides.txt file indeed show a distribution that is strongly skewed to the right. Common **preprocessing** steps therefore include $\log_2$-transformation to render the intensities more symmetric, and normalization to reduce systematic technical variation while retaining the underlying biological signal [79]. Other steps might include removal of common contaminant proteins (such as keratin from the operator's skin and hair, or leftover trypsin from digestion) [80] or bad quality peptides from the list of identified proteins [71].

Below, we illustrate how the robust peptide-based linear model from the MSqRob framework can be incorporated in a state-of-the-art label-free proteomics data analysis workflow. In a typical shotgun proteomics experiment, one would like to estimate the average $\log_2$ intensity per treatment for each protein. However, one also wants to correct for effects of the peptide sequence (which can be rather large, as explained in the historical background), pseudo-replication at the level of biological and technical repeats (i.e. MS runs) and other potential blocking factors. For each protein, a statistical peptide-based model is constructed in which one models all observed $\log_2$-transformed peptide intensities as a function of the effects in our model. A typical peptide-based model is formulated below:

$$y_{ijklmn} = \beta_{ij}^{treat} + \beta_{ik}^{pep} + \beta_{il}^{biorep} + \beta_{im}^{techrep} + \varepsilon_{ijklmn},$$

with $y_{ijklmn}$ the *n*th **preprocessed peptide intensity** for the *i*th protein, *j* is the index for treatment (*treat*), *k* the index for peptide sequence (*pep*), *l* the index for biological repeat (*biorep*) and the *m* the index for technical repeat (*techrep*). $\varepsilon_{ijklmn}$ is a normally distributed error term with mean zero and protein-specific variance $\sigma_i^2$. $y_{ijklmn}$ is also referred to as the **response variable**; and *treat*, *pep*, *biorep* and *techrep* as the **predictor variables**. The $\beta$'s are the **effects** of each predictor on the peptide intensities of the *i*th protein. More information about linear regression models can be found in Altman and Krzywinski [81].

When working with the MSqRob package, one needs to discriminate between **fixed** and **random effects** [82], as MSqRob handles them differently. Fixed effects are those effects for which all levels of interest are included in the experiment. They are generally controlled for by the experimenter and are typically the effects of interest. Examples include genotype (when comparing specific genetic constitutions), treatment, gender (only two levels), … Random effects are those effects for which not all levels are included in the experiment and the levels that are included can be considered to be drawn at random from a broader, near-infinite population. Experimenters are generally not interested in the observed random effect sizes, but they can be used to address issues with pseudo-replication, i.e. we merely incorporate them so that the covariance is correctly accounted for as to enable valid inference on the fixed effects of interest. Random effects are never under the control of the experimenter. Examples of random effects are MS run, biological replicate, technical replicate, animal effect, patient effect, etc. Note that the effect of the biological repeat (subjects/animals) only has to be incorporated if multiple observations are available per repeat. Sometimes, the number of observed levels also determines whether an effect is incorporated as fixed or random. E.g. if one performs an experiment with two different cell types, there are not enough levels to estimate the random effect variance, so it should be included as a fixed effect and the experimenter can only draw conclusions on the two specific cell types studied. In many cases blocking factors, such as effects of HPLC column or instrument, are also considered as fixed. They often have a limited number of levels and the variability between blocks can be factored out of the analysis in good experimental designs, i.e. when all effects of interest are included within each block. MSqRob also exploits the link between mixed models and ridge regression, which puts a penalty on the size of the fixed effects, preventing overfitting. The peptide effects often overwhelm the remaining effects in the experiment and specifying the sequence effect as a separate random effect allows the remaining fixed effects of interested to be penalized independently of the peptide effect.

Upon fitting a linear regression model, contrasts of the model parameters are assessed in statistical tests to answer the research question; e.g. one could test whether there is on average a difference between the effects of two treatments. Since the effects are modeled on a log-scale, differences can be interpreted in terms of $\log_2$ fold changes. Another option is to perform an ANOVA test to assess multiple contrasts simultaneously or the omnibus null hypothesis that none of the treatments have an effect.

Since we infer on the research question for each protein, it is necessary to correct for multiple testing [83], [84]. In high-throughput experiments we generally use the false discovery rate (FDR) for this purpose. Researchers often tolerate a few false positives in their top hits, as long as there are not too many. Controlling the FDR at 5% means that one expects on average 5% false positive proteins amongst all proteins that are returned as differentially abundant. In MSqRob, we correct for multiple testing using the Benjamini-Hochberg FDR procedure [85].

## How is MSqRob used in research?

MSqRob can be used in two ways: either as an R package or with the “Shiny” graphical user interface. Info on how to use the latest version of MSqRob in R can be found in the MSqRob vignette or in the installation instructions on the MSqRob github repository. MSqRob offers custom functions for importing data, preprocessing data, fitting models and testing research hypotheses (“statistical contrasts”). As long as peptide-level data can be provided in either long or wide tabular format, MSqRob can be used after searching the data with any search engine.

As MaxQuant is one of the most popular free quantitative proteomics software packages, we developed a graphical user interface for statistical analysis of differential protein abundance based on MaxQuant output. It allows to (1) directly import MaxQuant search results, (2) preprocess and visualize the data, and (3) save the output to Excel without any programming knowledge required. Moreover, MSqRob is capable of handling virtually any experimental design.

Our MSqRob Shiny App has three different tabs: an input, a preprocessing and a results tab. In the input tab, the user provides the name of the project, the location where the output needs to be saved, MaxQuant's peptides.txt file and an experimental annotation file. In the preprocessing tab, options are provided to $\log_2$-transform peptide intensities, normalize intensities, remove overlapping protein groups, remove contaminants and reverse sequences, remove all proteins that are only identified by modified peptides and remove all peptides that are identified by less than a specified number in the dataset. Its right panel shows diagnostic plots that can be used to evaluate the preprocessing step. Ultimately, the quantification tab allows the user to select the grouping factor, remove superfluous columns, select fixed and random effects and specify contrasts. When pressing the “Go” button, MSqRob will execute the analysis. After the analysis, the right panel of the quantification tab will show a volcano plot in which proteins can be selected for further inspection with a detail plot. This panel will also show the results table. The results table can be saved automatically to allow further inspection and visualization.

# Case studies

## Prerequisites

R [86] and RStudio [87] have to be installed on a computer. MSqRob can be freely downloaded from https://github.com/statOmics/MSqRob. Installation instructions and up-to-date guidelines are provided in the README.md file on the website.

MSqRob is an R package with a Shiny App that provides a graphical user interface to MSqRob for MaxQuant data. In the tutorial we focus on hands-on examples in the MSqRob Shiny App. The examples can also be coded in plain R, which can be useful for incorporating MSqRob in data analysis pipelines. R-markdown files with R Code and instructions are also provided for the examples at https://github.com/statOmics/MSqRob/blob/master/vignettes/MSqRob.Rmd.

Upon installation, the Shiny App can be launched by copy-pasting the following command in the command window of RStudio:

```
shiny::runApp(system.file('App-MSqRob', package = 'MSqRob'))
```

Here, we provide step-by-step tutorials for two case studies with the MSqRob Shiny application. Our first example is a case study based on the experiment of Ramond et al. [88]. We use a subset of the experiment with a simple wild-type vs. knock-out design. It is a design with pseudo-replication at different levels. Our second example consists of a spike-in study of the Clinical Proteomic Technology Assessment for Cancer Network (CPTAC) in which 48 human proteins were spiked in five different concentrations in a yeast background proteome. Here, the ground truth is known [89] and the experiment is set up as a randomized complete block design. We have already used this particular study to evaluate the performance of our method [74].

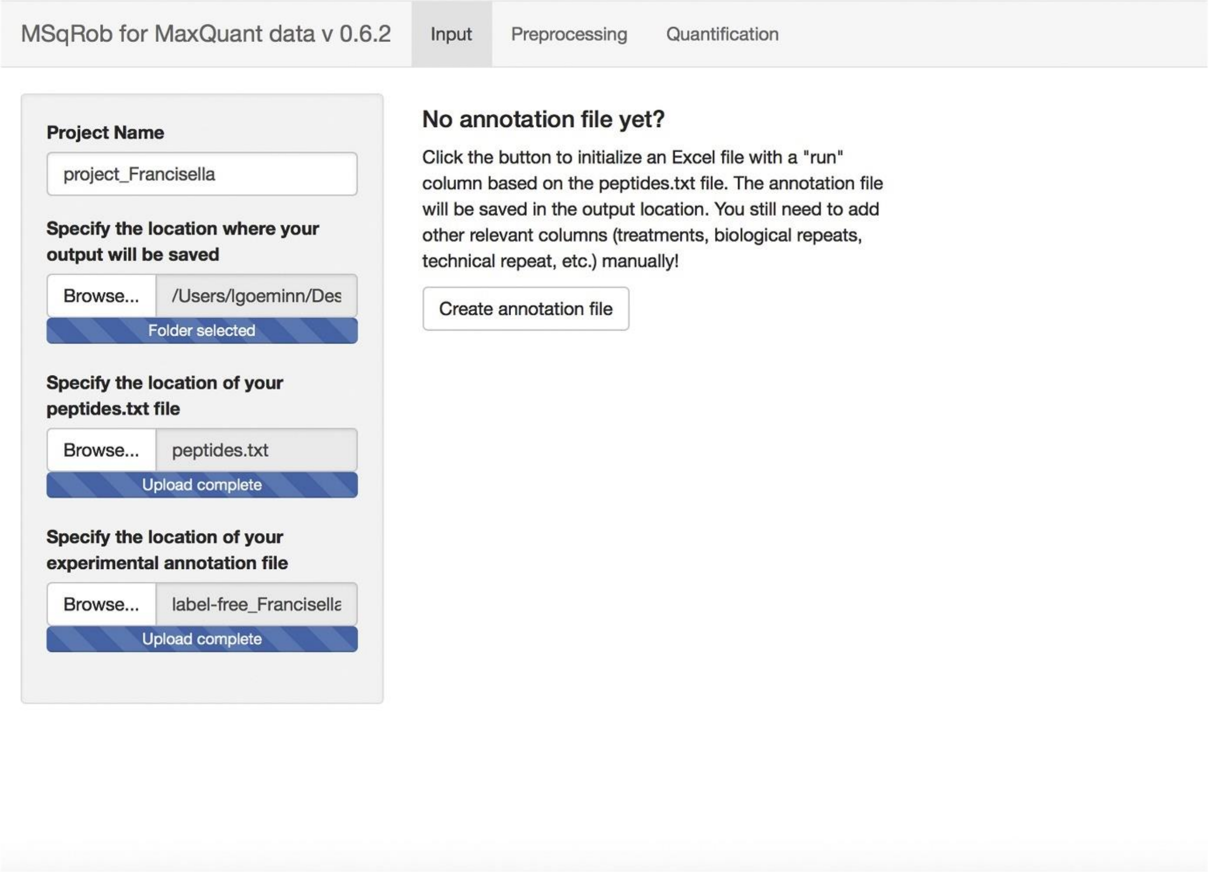
## The *Francisella* example

### *Experimental set-up*

The study on the facultative pathogen *Francisella* tularensis was conceived by Ramond et al. [88]. *F. tularensis* enters the cells of its host by phagocytosis. The authors showed that *F. tularensis* must import arginine from the host cell via a novel arginine transporter, ArgP, in order to efficiently escape from the phagosome and reach the cytosolic compartment, where it can actively multiply. In their study, they compared the proteome of wild type *F. tularensis* (WT) to ArgP-gene deleted *F. tularensis* (knock-out, KO). For this experiment, bacterial cultures were grown in biological triplicate and each sample was run three times on a nanoRSLC-Q Exactive PLUS instrument. Hence, pseudo-replication occurs on different levels of the experiment, i.e. multiple peptides for the same protein in each MS-run (technical repeat) and 3 technical repeats for each biological repeat. The data were searched with MaxQuant version 1.4.1.2. Below, we give an overview on how to process the data with the MSqRob Shiny App.

***The input tab (Fig. 3)***

First, we choose an appropriate **name** for the **project**. This name, appended with a timestamp, will be used to generate an output folder for the MSqRob model and results. Here, we use the name “project_Francisella”. Select an appropriate **file location** where the MSqRob output should be saved by clicking on “Browse…”. Next, upload the **peptides.txt file**, which contains the MaxQuant peptide-level intensities that are found by default in the “path_to_raw_files/combined/txt/” folder from the MaxQuant output, with “path_to_raw_files” the folder where raw files were saved.



Project Name: project_Francisella

Specify the location where your output will be saved: Browse... /Users/lgoeminn/Des
Folder selected

Specify the location of your peptides.txt file: Browse... peptides.txt
Upload complete

Specify the location of your experimental annotation file: Browse... label-free_Francisella
Upload complete

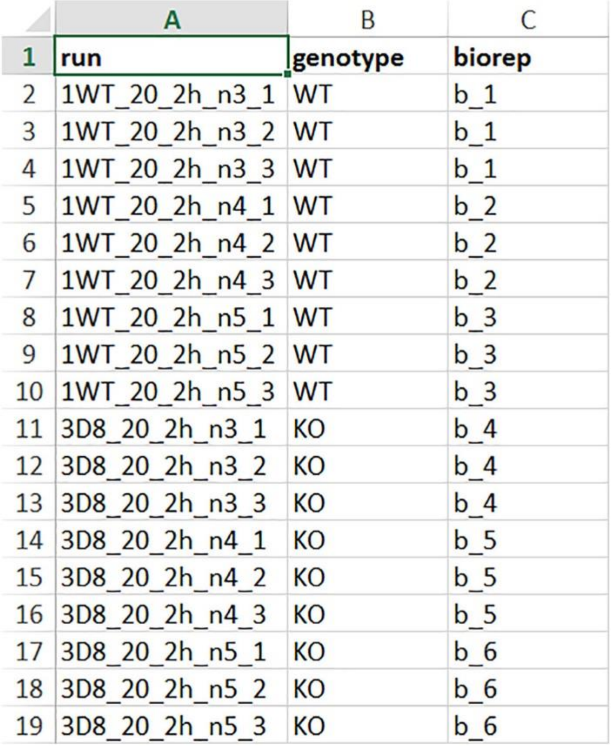
No annotation file yet?

Click the button to initialize an Excel file with a "run" column based on the peptides.txt file. The annotation file will be saved in the output location. You still need to add other relevant columns (treatments, biological repeats, technical repeat, etc.) manually!

**Figure 3.** Overview of MSqRob's input tab.

Similarly, upload the **experimental annotation file**. This file should be a tab-delimited file or an Office Open XML spreadsheet file (“.xlsx” file). If needed, this file can be made based on Fig. 4. If the file location was already specified and the peptides.txt file was uploaded, one can generate the “run” column of this file automatically by clicking the “Create annotation file” button. The other columns need to be filled in manually based on the experimental design.

Alternatively, one can download the file from https://github.com/statOmics/MSqRobData/blob/master/inst/extdata/Francisella/label-free_Francisella_annotation.xlsx. One column (the “run” column in Fig. 4) of the experimental annotation file should contain the names of the MS runs; i.e. the names given in the “experiment names” column when searching the data with MaxQuant. These names should be unique. Other columns indicate other variables of interest related to the design that can affect protein expression; e.g. genotype: WT vs. KO and biological repeats (“biorep”).



|  | A | B | C |
|---|---|---|---|
| 1 | **run** | **genotype** | **biorep** |
| 2 | 1WT_20_2h_n3_1 | WT | b_1 |
| 3 | 1WT_20_2h_n3_2 | WT | b_1 |
| 4 | 1WT_20_2h_n3_3 | WT | b_1 |
| 5 | 1WT_20_2h_n4_1 | WT | b_2 |
| 6 | 1WT_20_2h_n4_2 | WT | b_2 |
| 7 | 1WT_20_2h_n4_3 | WT | b_2 |
| 8 | 1WT_20_2h_n5_1 | WT | b_3 |
| 9 | 1WT_20_2h_n5_2 | WT | b_3 |
| 10 | 1WT_20_2h_n5_3 | WT | b_3 |
| 11 | 3D8_20_2h_n3_1 | KO | b_4 |
| 12 | 3D8_20_2h_n3_2 | KO | b_4 |
| 13 | 3D8_20_2h_n3_3 | KO | b_4 |
| 14 | 3D8_20_2h_n4_1 | KO | b_5 |
| 15 | 3D8_20_2h_n4_2 | KO | b_5 |
| 16 | 3D8_20_2h_n4_3 | KO | b_5 |
| 17 | 3D8_20_2h_n5_1 | KO | b_6 |
| 18 | 3D8_20_2h_n5_2 | KO | b_6 |
| 19 | 3D8_20_2h_n5_3 | KO | b_6 |

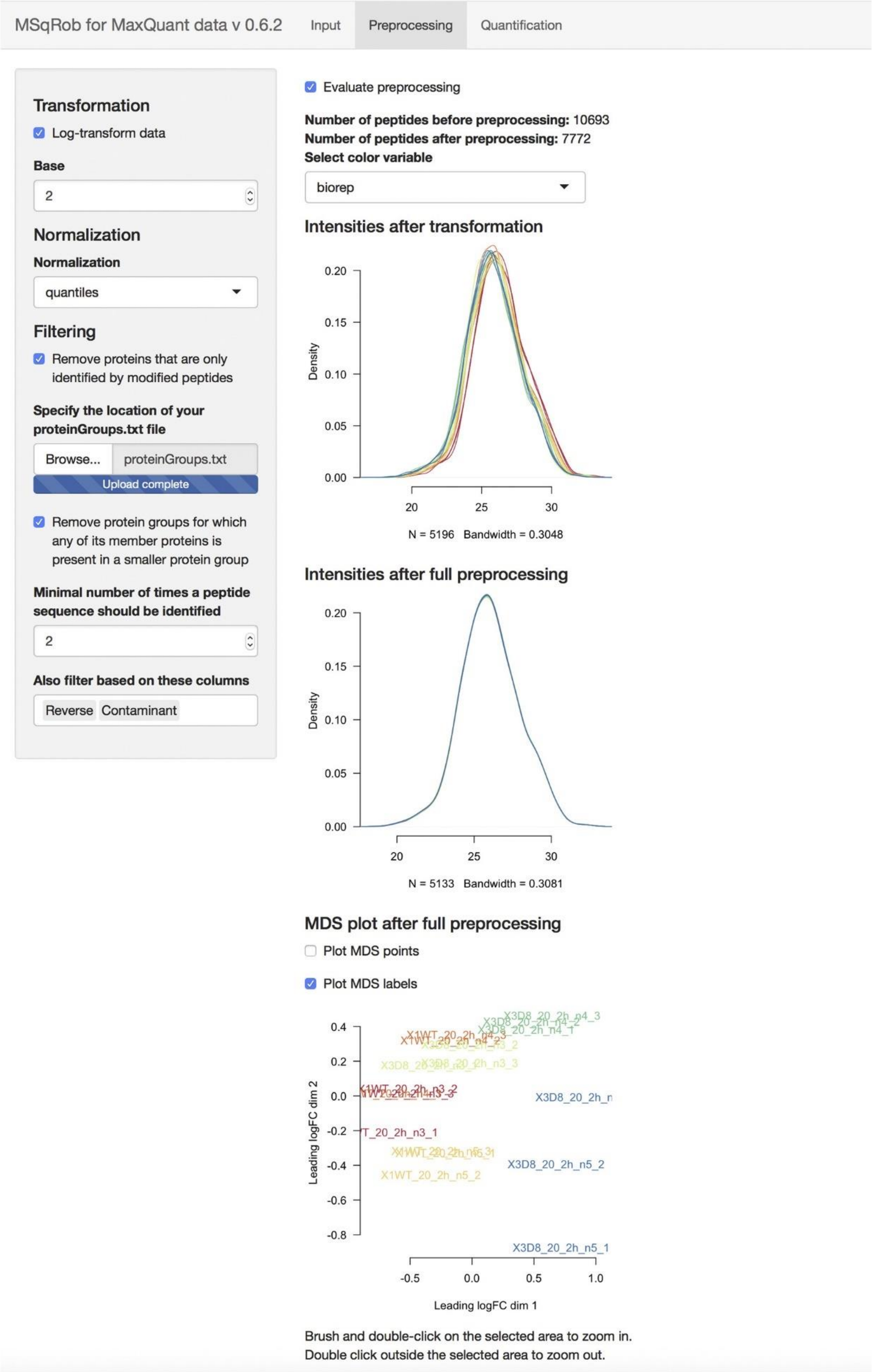
**Figure 4.** Experimental annotation file for the Francisella dataset.

At this stage, everything is set for preprocessing and data exploration, which are implemented in the preprocessing tab.

***The preprocessing tab (Fig. 5)***

*Left panel*

The preprocessing tab features different preprocessing options, many of which can be safely left at their default state.



Transformation

☑ Log-transform data

Base

2

Normalization

Normalization

quantiles

Filtering

☑ Remove proteins that are only identified by modified peptides

Specify the location of your proteinGroups.txt file

Browse... proteinGroups.txt

Upload complete

☑ Remove protein groups for which any of its member proteins is present in a smaller protein group

Minimal number of times a peptide sequence should be identified

2

Also filter based on these columns

Reverse Contaminant

☑ Evaluate preprocessing

**Number of peptides before preprocessing:** 10693

**Number of peptides after preprocessing:** 7772

**Select color variable**

biorep

Intensities after transformation

Intensities after full preprocessing

MDS plot after full preprocessing

□ Plot MDS points

☑ Plot MDS labels

Brush and double-click on the selected area to zoom in.
Double click outside the selected area to zoom out.

**Figure 5.** Overview of MSqRob's preprocessing tab.

MS-based proteomic intensity distributions are nearly always strongly skewed to the right. Therefore, a log-**transformation** is highly recommended. We suggest to log-transform the data with **base** 2. This has the added advantage that the model estimates will be interpreted as $\log_2$ FC. For the remainder of this work, we assume that intensities have been $\log_2$-transformed.

We provide different **normalization** approaches. As a default, we suggest quantile normalization [79], [90]. Quantile normalization imposes the same empirical intensity distribution on all runs. More information on other normalization methods that are implemented can be found in the documentation of the 'normalise' function in the R package MSnbase [91]. The effect of quantile normalization on the distribution of the $\log_2$-transformed peptide intensities is shown in Fig. 6.

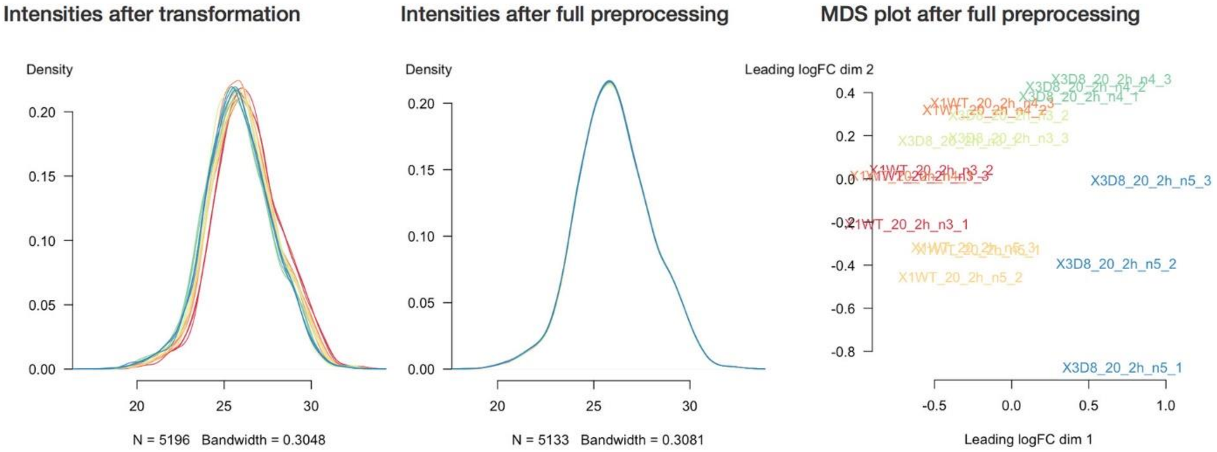


**Figure 6.** Overview of the $\log_2$-transformed peptide intensities for the Francisella dataset before and after preprocessing. Left: $\log_2$-transformed peptide intensities before preprocessing, center: $\log_2$-transformed peptide intensities after preprocessing. Note that the densities are forced onto the same distribution. Right: MDS plot that clusters similar MS runs together.

The option **"Remove proteins that are only identified by modified peptides"** allows for removing proteins that are only identified by peptides that carry one or more modified amino acids. Identification of such peptides in the background of non-modified peptides is often less reliable, and proteins only identified by such peptides are therefore removed in a typical MaxQuant-Perseus workflow. We offer the option to do a similar filtering in MSqRob. The MaxQuant's proteinGroups.txt file is needed for this purpose and can be found in the "combined/txt/" folder.

Razor peptides are peptides that cannot be uniquely attributed to a single protein or protein group. As we are uncertain from which protein group these peptides originate and their intensities might even be a combined value from multiple protein groups, we opt to remove these peptides by default. The option **"Remove protein groups for which any of its member proteins is present in a smaller protein group"** deals with peptides that are shared between protein groups. This option removes all peptides in protein groups for which any of its peptides map to a protein that is also present in another smaller protein group.

"Minimal number of times a peptide sequence should be identified" indicates a threshold *T* for how many times a certain peptide sequence should be present in the data before being retained in the final analysis. Peptides that are identified at least *T* times are retained; other peptides are removed from the data. This value defaults to 2 and there is a very practical reason for this. Indeed, we need a parameter in the model for each peptide sequence. Adding a parameter for a single observation leads to perfect confounding in the model as there is no way to discern between the peptide-specific effect and the other effects for this observation. Note that this is not the same as applying the so-called "two-peptide rule" [92]. A protein identified by only one peptide can contribute to the estimation provided that the peptide is identified in multiple samples, say $t$ with $t \geq T$.

One can further filter out reverse sequences and potential contaminants, made possible by providing the column names of the peptides.txt file that indicate these sequences in the **“Also filter based on these columns”** field.

*Right panel*

In the right panel, the number of peptides before any kind of preprocessing is done, and a plot of the densities of the (log-transformed) peptide intensities in each MS run are displayed. For the *Francisella* dataset, there were 10,693 identified peptides before preprocessing.

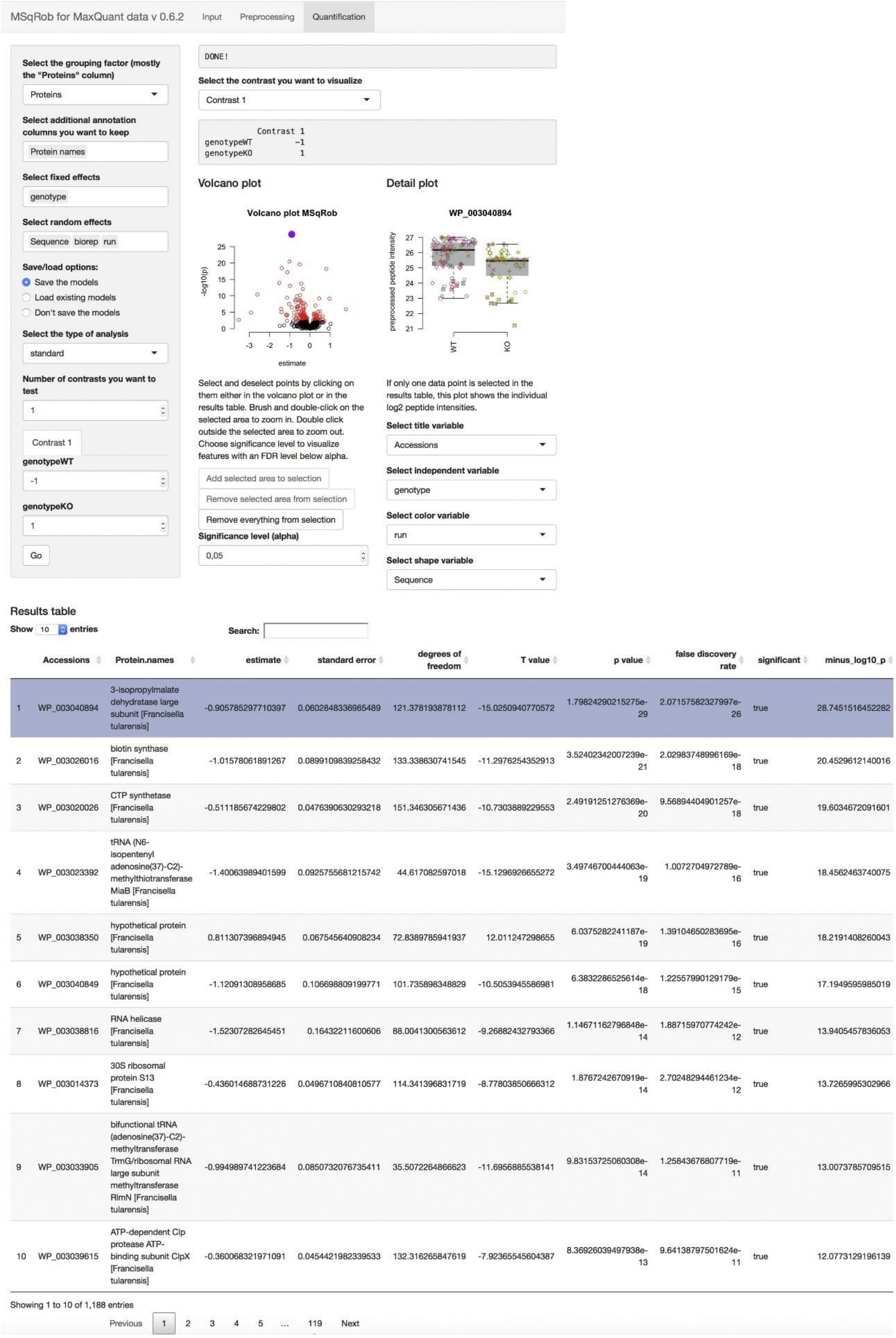
The effect of preprocessing can be assessed by ticking the **“Evaluate preprocessing”** box. A plot will be generated that shows the densities of the (log-transformed) peptide intensities after full preprocessing (i.e. after normalization and filtering) (Fig. 6). A multidimensional scaling (MDS) plot will also be produced, which shows a dot for each MS run such that the distance between two dots is equal to the root-mean-square deviation for the top 500 peptides that distinguish the two corresponding runs. Options are provided to show only dots, only labels or both. It is also possible to zoom in on a particular part of the plot by dragging the mouse to select a particular area on the plot and then double-click to zoom in.

One can color the density lines and MDS points by any factor provided in the experimental annotation file. Upon filtering, $\log_2$-transformation and normalization, 7772 peptides remained in the dataset.

***The quantification tab (Fig. 7)***

*Left panel*

**“Select the grouping factor (mostly the “Proteins” column)”** allows selecting on which level the statistical inference is performed. Here, we were interested in proteins of which the abundance differed between the two genotypes. We thus selected the “Proteins” column. **“Select additional annotation columns you want to keep”** allows retaining extra annotation columns that one might have added to the peptides.txt file. Here, we selected “Protein names” and “GI number”.



Select the grouping factor (mostly the "Proteins" column): Proteins

Select additional annotation columns you want to keep: Protein names

Select fixed effects: genotype

Select random effects: Sequence biorep run

Save/load options:

- ● Save the models
- ○ Load existing models
- ○ Don't save the models

Select the type of analysis: standard

Number of contrasts you want to test: 1

Contrast 1

genotypeWT: -1

genotypeKO: 1

```
DONE!
```

Select the contrast you want to visualize: Contrast 1

```
           Contrast 1
genotypeWT         -1
genotypeKO          1
```

Select and deselect points by clicking on them either in the volcano plot or in the results table. Brush and double-click on the selected area to zoom in. Double click outside the selected area to zoom out. Choose significance level to visualize features with an FDR level below alpha.

Significance level (alpha): 0,05

If only one data point is selected in the results table, this plot shows the individual log2 peptide intensities.

Select title variable: Accessions

Select independent variable: genotype

Select color variable: run

Select shape variable: Sequence

Results table

Show 10 entries

Search:

| | Accessions | Protein.names | estimate | standard error | degrees of freedom | T value | p value | false discovery rate | significant | minus_log10_p |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | WP_003040894 | 3-isopropylmalate dehydratase large subunit [Francisella tularensis] | -0.905785297710397 | 0.0602848336965489 | 121.378193878112 | -15.0250940770572 | 1.79824290215275e-29 | 2.07157582327997e-26 | true | 28.7451516452282 |
| 2 | WP_003026016 | biotin synthase [Francisella tularensis] | -1.01578061891267 | 0.0899109839258432 | 133.338630741545 | -11.2976254352913 | 3.52402342007239e-21 | 2.02983748996169e-18 | true | 20.4529612140016 |
| 3 | WP_003020026 | CTP synthetase [Francisella tularensis] | -0.511185674229802 | 0.0476390630293218 | 151.346305671436 | -10.7303889229553 | 2.49191251276369e-20 | 9.56894404901257e-18 | true | 19.6034672091601 |
| 4 | WP_003023392 | tRNA (N6-isopentenyl adenosine(37)-C2)-methylthiotransferase MiaB [Francisella tularensis] | -1.40063989401599 | 0.0925755681215742 | 44.617082597018 | -15.1296926655272 | 3.49746700444063e-19 | 1.0072704972789e-16 | true | 18.4562463740075 |
| 5 | WP_003038350 | hypothetical protein [Francisella tularensis] | 0.811307396894945 | 0.067545640908234 | 72.8389785941937 | 12.011247298655 | 6.0375282241187e-19 | 1.39104650283695e-16 | true | 18.2191408260043 |
| 6 | WP_003040849 | hypothetical protein [Francisella tularensis] | -1.12091308958685 | 0.106698809199771 | 101.735898348829 | -10.5053945586981 | 6.3832286525614e-18 | 1.22557990129179e-15 | true | 17.1949595985019 |
| 7 | WP_003038816 | RNA helicase [Francisella tularensis] | -1.52307282645451 | 0.16432211600606 | 88.0041300563612 | -9.26882432793366 | 1.14671162796848e-14 | 1.88715970774242e-12 | true | 13.9405457836053 |
| 8 | WP_003014373 | 30S ribosomal protein S13 [Francisella tularensis] | -0.436014688731226 | 0.0496710840810577 | 114.341396831719 | -8.77803850666312 | 1.8767242670919e-14 | 2.70248294461234e-12 | true | 13.7265995302966 |
| 9 | WP_003033905 | bifunctional tRNA (adenosine(37)-C2)-methyltransferase TrmG/ribosomal RNA large subunit methyltransferase RlmN [Francisella tularensis] | -0.994989741223684 | 0.0850732076735411 | 35.5072264866623 | -11.6956885538141 | 9.83153725060308e-14 | 1.25843676807719e-11 | true | 13.0073785709515 |
| 10 | WP_003039615 | ATP-dependent Clp protease ATP-binding subunit ClpX [Francisella tularensis] | -0.360068321971091 | 0.0454421982339533 | 132.316265847619 | -7.92365545604387 | 8.36926039497938e-13 | 9.64138797501624e-11 | true | 12.0773129196139 |

Showing 1 to 10 of 1,188 entries

**Figure 7.** Overview of MSqRob's quantification tab.

Next, the **fixed effects** need to be specified. Fixed effects are effects that remain constant when repeating the experiment. For factor variables (e.g. genotype), the number of levels are typically small. “genotype” should be entered as a fixed effect, as there are only 2 genotypes in our study. Effects of interest are nearly always fixed effects. On the contrary, “Sequence”, “biorep” and “run” are added as **random effects**. The effect of a single biological repeat will differ each time one would re-perform an experiment. The biological repeat also has to be included in the model because peptide intensities from a protein from the same biological repeat are more similar than those from the same protein across biological repeats. Similarly, each MS run will be different and peptides from the same protein in the same run are correlated because they originate from the same protein pool. Hence, the pseudo-replication of peptides within technical repeats as well as the technical repeats within each biological repeat will be properly addressed. Assigning “Sequence” as a random effect is debatable, but we noticed that the sequence effect overwhelms other effects in typical proteomics experiments. MSqRob also exploits the link between ridge regression and mixed models [74]. Ridge regression is implemented to prevent overfitting. Therefore, we strongly suggest specifying the “Sequence” effect as a random effect, which will allow penalizing this effect separately from the remaining fixed effects.

With **“Save/load options”**, there are three options:

1. “Save the models” will generate a file with an “.rDatas” extension that contains R objects with the data and the fitted models. It is useful to store these objects as they enable the user to upload and redo the statistical inference without having to perform the time-consuming preprocessing and model fitting steps.

2. “Load existing models” allows the user to upload an rDatas object from a previous analysis. Note that all input except the type of analysis and the contrast options will become disabled as the model is already fitted to the data. A new rDatas object will also be created with the output. This option is also useful for evaluating the output of MSqRob upon running it in bash mode.

3. “Don't save the models”: no rDatas object will be stored.

**“Number of contrasts you want to test”** indicates how many contrasts (research hypotheses) one would like to test. For the *Francisella* dataset, we were only interested in the difference between wild-type and knock-out strains, therefore we performed statistical inference on the average difference in $\log_2$ protein intensity between both genotypes. This difference corresponds to a $\log_2$ FC. We specify this contrast as by typing “− 1” under genoWT and “1” under genoKO.

Check all settings and press the “Go” button in the left panel of the output tab.

*Right panel*

When the analysis is finished, MSqRob prints “DONE!” at the top of the right panel. In this case study we only evaluated one contrast (KO vs. WT). If multiple contrasts are specified, one can select the contrast one would like to explore further. The **“Volcano plot”** shows $-\log_{10}$(p-values) as a function of the “estimate” (i.e., here the $\log_2$ FC between KO and WT for the *Francisella* example). One can select an area on this plot using the computer mouse and double clicking zooms in on this area. Upon selecting such an area, one can add all points in the area to a selection or remove all points in this area from a selection using their respective buttons. By clicking on a dot, one selects/deselects it. When only one protein is selected, a **“Detail plot”** is made for this protein, which shows the preprocessed peptide intensities as a function of a predictor variable from the model. Boxplots show the median preprocessed peptide intensity as a thick black line, the box itself comprises the interquartile range (IQR) and whiskers extend to the most extreme data point that lies within 1.5 times the IQR on each side [93]. Each peptide intensity in the Detail plot can be given a color and a shape value according to any model parameter. Fig. 8 shows a detail plot for the most significant protein in the study, WP_003040894 or 3-isopropylmalate dehydratase large subunit, an enzyme required for the biosynthesis of leucine. Note that all identified enzymes required for the synthesis of branched chain amino acids were found either unchanged or downregulated in the ArgP mutant. Here, we specified “genotype” as independent variable, “run” as color variable and “Sequence” as shape variable.

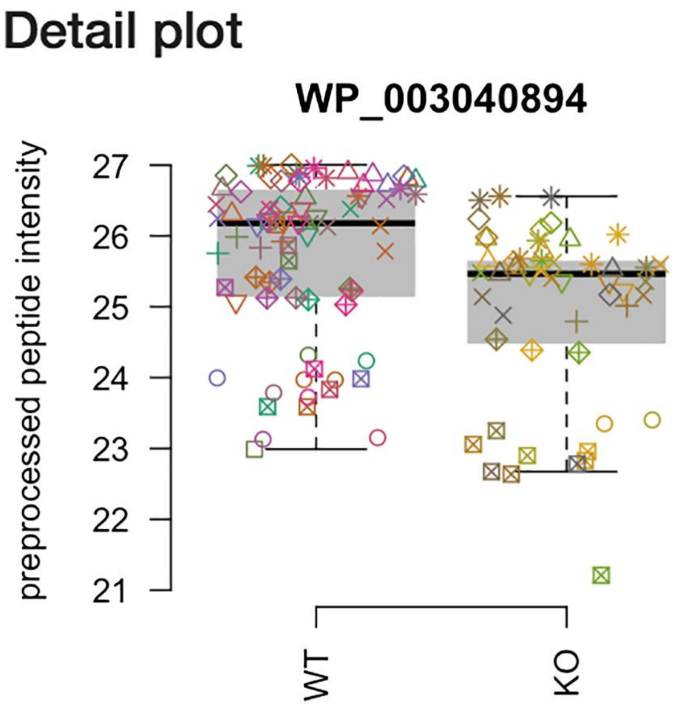


**Figure 8.** Example of a detail plot for the most significant protein in the study performed: 3-isopropylmalate dehydratase large subunit.

One may also specify the **“Significance level (alpha)”**. Its default value is at 5%, but it can be changed on the fly. Proteins with a false discovery rate (FDR) below α will be colored red in the Volcano plot if unselected, and purple if selected. Proteins with an FDR above α will be colored black if unselected, and grey if selected. In the *Francisella* case study, we found 162 significant proteins at a 5% FDR threshold. 154 of those overlap with the proteins reported in Goeminne et al. [74]. This difference is due to subtle changes in our algorithm (e.g. all fixed effects except peptide sequence now all get the same shrinkage penalty). The 8 new significant proteins are proteins for which our old implementation could not estimate a fold change, while the 5 proteins that are not flagged anymore in our new implementation have an FDR value that is close to the 5% cut-off.

The **“Results table”** shows all proteins, by default sorted from smallest to largest p value. When selecting/deselecting a row in the table, the corresponding dot in the Volcano plot is also (de)selected and vice versa. When zoomed in on the Volcano plot, the Results table only shows the proteins corresponding to the dots in the plot window. The **Search** box allows searching for particular proteins in the table. When only one protein is selected, the detail plot is also displayed.

Note that in the file location one provided in the input tab, a folder is created, which is named “project_Francisella_*[date and time of analysis]*”. In this folder, one finds the project_Francisella_models.rDatas file, which contains the data and the fitted model object as discussed above. The “project_Francisella_results.xlsx” contains the same information as the “Results” table. The first column is the protein accession. **“Protein.names”** and **“GI.number”** are the columns, which were indicated as “additional columns we want to keep”. **“estimate”** is the estimate of the contrast, which here is the $\log_2$ FC between the proteomes of wild-type and knock-out *Francisella tularensis*. **“se”** is the standard error on the contrast, **“df”** indicates the degrees of freedom, **“Tval”** is the T value, **“pval”** is the p value, **“qval”** is the q value, i.e. the minimal FDR level at which this protein will be called significant. **“signif”** indicates whether the protein is significant at the default 5% FDR threshold.

The same analysis can also be performed in bash mode. Details are given at https://github.com/statOmics/MSqRob.

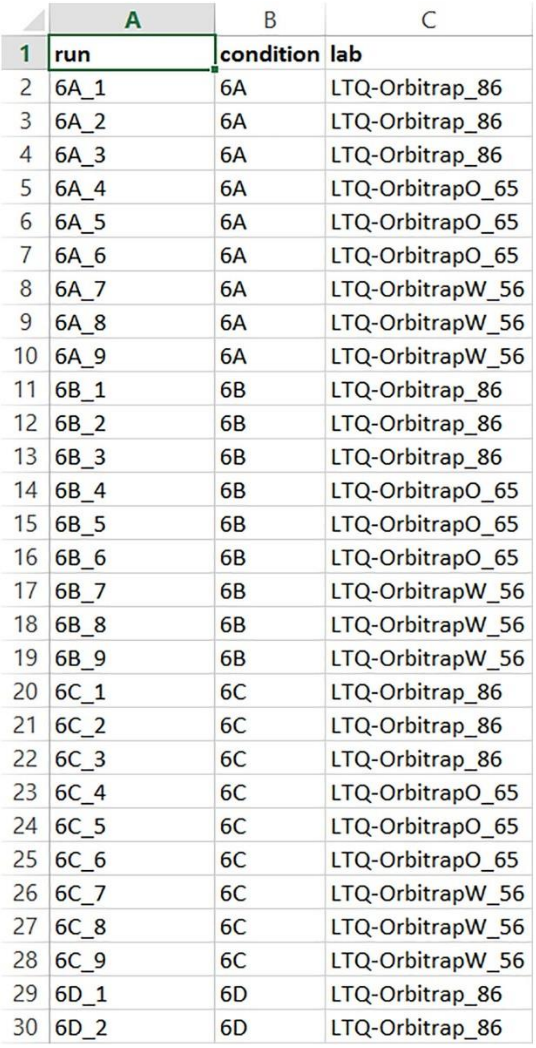
## The CPTAC example

The 6th study of the Clinical Proteomic Technology Assessment for Cancer (CPTAC) is an experiment in which the authors spiked the Sigma Universal Protein Standard mixture 1 (UPS1) containing 48 different human proteins in a protein background of 60 ng/μL *Saccharomyces cerevisiae* strain BY4741 (MATa, leu2Δ0, met15Δ0, ura3Δ0, his3Δ1). Five different spike-in concentrations were used: 6A (0.25 fmol UPS1 proteins/μL), 6B (0.74 fmol UPS1 proteins/μL), 6C (2.22 fmol UPS1 proteins/μL), 6D (6.67 fmol UPS1 proteins/μL) and 6E (20 fmol UPS1 proteins/μL) [89]. The raw data files can be downloaded from https://cptac-data-portal.georgetown.edu/cptac/public?scope=Phase+I (Study 6). We limited ourselves to the data of LTQ-Orbitrap at site 86, LTQ-Orbitrap O at site 65 and LTQ-Orbitrap W at site 56. The data were searched with MaxQuant version 1.5.2.8, and detailed search settings were described in Goeminne et al. [74]. The experiment is conceived as a randomized complete block design with lab as a blocking factor. For every lab, 3 replicates are available for each concentration.

At high spike-in concentrations of human proteins, especially in conditions 6D and 6E, ionization suppression of yeast proteins has been reported [54], [73], [74]. Therefore, we focus on differences between condition 6B–6A, 6C–6A, and 6C–6B.

### *The input tab*

Again, an appropriate **name** is chosen for the **project**. Here, use “project_CPTAC”, select the **file location** where the output has to be saved. Next, the location of the experimental annotation file and the peptides.txt file is specified, and MaxQuant's peptides.txt file is imported. An example of the experimental annotation for the CPTAC dataset is given in Fig. 9. This file can be downloaded from https://github.com/statOmics/MSqRobData/blob/master/inst/extdata/CPTAC/label-free_CPTAC_annotation.xlsx.



| | A | B | C |
|---|---|---|---|
| 1 | run | condition | lab |
| 2 | 6A_1 | 6A | LTQ-Orbitrap_86 |
| 3 | 6A_2 | 6A | LTQ-Orbitrap_86 |
| 4 | 6A_3 | 6A | LTQ-Orbitrap_86 |
| 5 | 6A_4 | 6A | LTQ-OrbitrapO_65 |
| 6 | 6A_5 | 6A | LTQ-OrbitrapO_65 |
| 7 | 6A_6 | 6A | LTQ-OrbitrapO_65 |
| 8 | 6A_7 | 6A | LTQ-OrbitrapW_56 |
| 9 | 6A_8 | 6A | LTQ-OrbitrapW_56 |
| 10 | 6A_9 | 6A | LTQ-OrbitrapW_56 |
| 11 | 6B_1 | 6B | LTQ-Orbitrap_86 |
| 12 | 6B_2 | 6B | LTQ-Orbitrap_86 |
| 13 | 6B_3 | 6B | LTQ-Orbitrap_86 |
| 14 | 6B_4 | 6B | LTQ-OrbitrapO_65 |
| 15 | 6B_5 | 6B | LTQ-OrbitrapO_65 |
| 16 | 6B_6 | 6B | LTQ-OrbitrapO_65 |
| 17 | 6B_7 | 6B | LTQ-OrbitrapW_56 |
| 18 | 6B_8 | 6B | LTQ-OrbitrapW_56 |
| 19 | 6B_9 | 6B | LTQ-OrbitrapW_56 |
| 20 | 6C_1 | 6C | LTQ-Orbitrap_86 |
| 21 | 6C_2 | 6C | LTQ-Orbitrap_86 |
| 22 | 6C_3 | 6C | LTQ-Orbitrap_86 |
| 23 | 6C_4 | 6C | LTQ-OrbitrapO_65 |
| 24 | 6C_5 | 6C | LTQ-OrbitrapO_65 |
| 25 | 6C_6 | 6C | LTQ-OrbitrapO_65 |
| 26 | 6C_7 | 6C | LTQ-OrbitrapW_56 |
| 27 | 6C_8 | 6C | LTQ-OrbitrapW_56 |
| 28 | 6C_9 | 6C | LTQ-OrbitrapW_56 |
| 29 | 6D_1 | 6D | LTQ-Orbitrap_86 |
| 30 | 6D_2 | 6D | LTQ-Orbitrap_86 |

**Figure 9.** Top 30 rows of the annotation file for the CPTAC dataset.

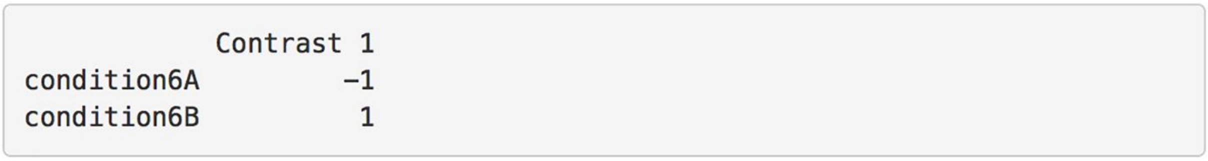
The preprocessing part is analogous as for the *Francisella* example.

### *The quantification tab*

We again grouped by "Proteins", but now there is no interest in additional columns. We selected "condition" as a fixed effect, because it is the main effect of interest and it has a fixed number of levels, being one for each spike-in concentration. The "lab" effect can be considered fixed, as it is a typical example of a so-called block effect. If one would redo the experiment, it will probably be in the same three labs, although it is also possible to argue for "lab" as a random effect (when one considers "lab" as a random draw from a huge number of possible labs). However, for the analysis of the treatment effect this should not matter as all treatment effects are observed within a lab and one can thus factor out the lab-to-lab variability from the analysis [77]. "Sequence" and "run" are again specified as random effects.

In this example, we assessed three contrasts of interest; thus, set **"Number of contrasts you want to test"** to 3. For the first contrast, set condition 6A to − 1 and condition 6B to 1, for the second contrast, set condition 6A to − 1 and condition 6C to 1 and for the third contrast, set condition 6B to − 1 and condition 6C to 1 for comparisons 6B–6A, 6C–6B and 6C–6A, respectively. Then press the **"Go"** button and wait for the analysis to complete.

Upon comparing condition 6B to condition 6A on the Volcano plot, we noticed that most hits have positive FC estimates. These red circles on the right of “0” are indeed the UPS1 spike-in proteins and their levels in condition 6B are higher than in condition 6A. There appear to be two false positive hits (red circle left of “0” and the selected purple circle in Fig. 10). The selected yeast protein sp | P53115 | INO80_YEAST exhibits a strong negative $\log_2$ FC estimate of − 2.09 on the Volcano plot (Fig. 10). Upon inspecting this protein in the Detail plot, we found that this protein was only identified by two different peptides with very different intensity patterns (NAPSEGVMASLLNVEK: square and VSTTPLLK: circle). The intensities of the former peptide remain basically unchanged over the different spike-in concentrations, while those of the latter show a clear upwards trend with increasing spike-in concentration. Based on its intensity pattern, this latter peptide is very likely an incorrectly annotated UPS1 peptide. Indeed, our model assumes that all peptides behave in a similar way when comparing over samples. Here, the effect of the VSTTPLLK peptide is on average lower in condition 6B compared to the rest of the dataset, pulling the estimated average $\log_2$-intensity in this condition down. This effect is not at play for condition 6A, as this peptide was not identified, and therefore, the difference between 6B and 6A will be strongly negative (− 2.09). This example clearly demonstrates the added value of using Detail plots, as these enable detecting aberrations in the data that would otherwise go unnoticed, preventing researchers from drawing wrong conclusions.



```
            Contrast 1
condition6A         -1
condition6B          1
```

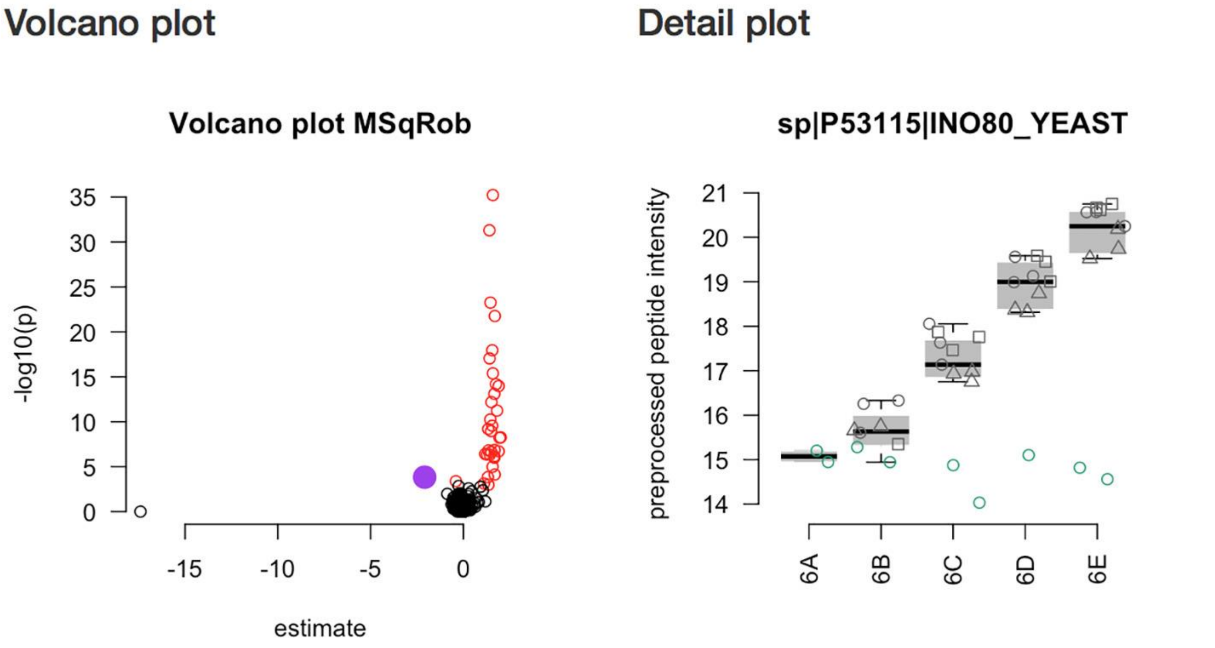


**Figure 10.** Use of the MSqRob output plots to find peculiar proteins. Protein sp | P53115 | INO80_YEAST has a log2 FC of − 2.09 and is identified by two different peptides, one of which is likely mis-annotated. In the Detail plot, points are colored by lab, while a different shape denotes a different peptide sequence.

## Current limitations and useful working limits

A major limitation of current proteomics workflows is the sequencing depth. The workflow we described here concerns so-called data-dependent acquisition (DDA). This means that identification is data-driven: only the most abundant peptide precursor ions are identified following $MS^2$. As a consequence, not all peptides in a sample are identified, which gives rise to missing values. Missingness in proteomics datasets is a combination of missingness completely at random (MCAR) (e.g. a misidentified peptide can either be identified by aligning its elution profile with an already identified peptide or this alignment can be missed) or not at random (MNAR) (e.g. more abundant peptides simply have a higher chance of getting fragmented and thus being identified) [94]. This is even exacerbated as the probability is also context-dependent: when co-eluting with many other highly intense peptides, a certain peptide will have a smaller chance of getting identified than if these other peptides would be absent or lower in numbers and/or abundance. Imputing missing peptide values was suggested in the proteomics literature, but imputation should always be used with caution. When nothing is known about the nature of the missing values, recent reviews suggest the use of MCAR imputation approaches based on local similarity, as these perform well on average [94]. It has to be noted, however, that the performance of an imputation approach is highly dataset-dependent [73], [94], [95]. Due to these peculiarities, we have chosen to omit imputation in the standard MSqRob workflow. Of course, when using non-imputed datasets, differential abundance cannot be estimated when all peptides are absent in all replicates of a particular condition. However, researchers have the option to impute peptide intensities before feeding these into MSqRob. Another solution to the presence-absence problem would be to perform an easy-to-implement spectral count approach to detect these proteins before continuing with a more sensitive intensity-based method [48]. So-called data-independent acquisition (DIA) workflows fragment all peptides, typically within a given m/z-window. With DIA, challenges lie in de-convoluting the highly complicated mixed spectra [96]. In this context missingness is due to the inability to resolve a spectrum but is expected to be less intensity-dependent.

Another major issue in proteomics bioinformatics is data standardization [97]. As MS-based proteomics becomes more and more affordable to a wider community of researchers, the number of customized and multidimensional experiments (i.e. experiments in which more than one protein property, such as abundance, modifications, turnover, localization, etc. is analyzed simultaneously) is expected to rise [98], [99]. Such experiments require customized workflows, however, many proteomics tools work with software-specific or even proprietary data formats. This makes it difficult to connect different tools in a customized workflow. Therefore, open data formats for storing proteomics data have been developed by HUPO. Examples of these are mzML for raw mass spectrometer output [100] and mzQuantML [101] for quantified peptides and proteins. Future adaptation of these formats will allow for more interconnectivity between applications and massively improve the feasibility of setting up custom workflows.

## Future developments

On a short term, we intend to adapt MSqRob to be able to handle DIA and isobaric labeling, and to enable the input of other search engines and open data formats.

A constant theme in improvements of mass spectrometry instruments has been their increase in analysis speed and proteome coverage. We expect this trend to continue, which could, in the long run, reduce or even eliminate intensity-dependent missingness. Faster machines will also allow biologists to analyze an increasing number of biological repeats, which will boost the power of their experiments and allow them to detect small, but sometimes very relevant perturbations with greater confidence. Thanks to such increasing coverage, each generation of machines allows us to dive deeper into the proteome than ever before. As machine duty cycles continue to increase, DIA and DDA are expected to come closer together as DIA windows will become smaller and smaller [102], while the analysis depth in DDA will continue to increase, so that one day, they might merge into a single technique that is capable of identifying all peptides in a sample. When that happens, the need to handle missing data in DDA will become obsolete.

## Acknowledgements

Part of this research was supported by IAP research network “StUDyS” grant no. P7/06 of the Belgian government (Belgian Science Policy) and the Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” of Ghent University. L.G. is supported by a Ph.D. grant from the Flanders Innovation & Entrepreneurship agency, Flanders (Agentschap Innoveren & Ondernemen – Vlaanderen) entitled ‘Differential proteomics at peptide, protein and module level’ (141573).

## Appendix A

Tutorial slide show (5MB)

## References

[1] A. Pandey, M. Mann. **Proteomics to study genes and genomes.** Nature, 405 (6788) (2000), pp. 837-846

[2] S. Hanke, H. Besir, D. Oesterhelt, M. Mann. **Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level.** J. Proteome Res., 7 (3) (2008), pp. 1118-1130

[3] P. Picotti, B. Bodenmiller, L.N. Mueller, B. Domon, R. Aebersold. **Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics.** Cell, 138 (4) (2009), pp. 795-806

[4] P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, R. Aebersold. **High-throughput generation of selected reaction-monitoring assays for proteins and proteomes.** Nat. Methods, 7 (1) (2010), pp. 43-46

[5] E. Ahrné, L. Molzahn, T. Glatter, A. Schmidt. **Critical assessment of proteome-wide label-free absolute abundance estimation strategies.** Proteomics, 13 (17) (2013), pp. 2567-2578

[6] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold. **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** Nat. Biotechnol., 17 (10) (1999), pp. 994-999

[7] Y. Oda, K. Huang, F.R. Cross, D. Cowburn, B.T. Chait. **Accurate quantitation of protein expression and site-specific phosphorylation.** Proc. Natl. Acad. Sci., 96 (12) (1999), pp. 6591-6596

[8] S.-E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann. **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** Mol. Cell. Proteomics, 1 (5) (2002), pp. 376-386

[9] S.-E. Ong, M. Mann. **A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC).** Nat. Protoc., 1 (6) (2007), pp. 2650-2660

[10] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D.J. Pappin. **Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents.** Mol. Cell. Proteomics, 3 (12) (2004), pp. 1154-1169

[11] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, C. Hamon. **Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.** Anal. Chem., 75 (8) (2003), pp. 1895-1904

[12] N. Rauniyar, J.R. Yates. **Isobaric labeling-based relative quantification in shotgun proteomics.** J. Proteome Res., 13 (12) (2014), pp. 5293-5309

[13] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster. **Quantitative mass spectrometry in proteomics: a critical review.** Anal. Bioanal. Chem., 389 (4) (2007), pp. 1017-1031

[14] V.J. Patel, K. Thalassinos, S.E. Slade, J.B. Connolly, A. Crombie, J.C. Murrell, J.H. Scrivens. **A comparison of labeling and label-free mass spectrometry-based proteomics approaches.** J. Proteome Res., 8 (7) (2009), pp. 3752-3759

[15] S. Tyanova, T. Temu, J. Cox. **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.** Nat. Protoc., 11 (12) (2016), pp. 2301-2319

[16] C.D. Kelstrup, R.R. Jersie-Christensen, T.S. Batth, T.N. Arrey, A. Kuehn, M. Kellmann, J.V. Olsen. **Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer.** J. Proteome Res., 13 (12) (2014), pp. 6187-6195

[17] S. Eliuk, A. Makarov. **Evolution of orbitrap mass spectrometry instrumentation.** Annu. Rev. Anal. Chem., 8 (1) (2015), pp. 61-80

[18] J.V. Olsen, S.-E. Ong, M. Mann. **Trypsin cleaves exclusively C-terminal to arginine and lysine residues.** Mol. Cell. Proteomics, 3 (6) (2004), pp. 608-614

[19] M. Wilm. **Principles of electrospray ionization.** Mol. Cell. Proteomics, 10 (7) (2011) (M111.009407)

[20] J. Mitchell Wells, S.A. McLuckey. **Collision-induced dissociation (CID) of peptides and proteins, Methods in Enzymology.** Academic Press (2005), pp. 148-185

[21] J.V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, M. Mann. **Higher-energy C-trap dissociation for peptide modification analysis.** Nat. Methods, 4 (9) (2007), pp. 709-712

[22] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold. **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** Anal. Chem., 74 (20) (2002), pp. 5383-5392

[23] J.K. Eng, A.L. McCormack, J.R. Yates. **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** J. Am. Soc. Mass Spectrom., 5 (11) (1994), pp. 976-989

[24] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell. **Probability-based protein identification by searching sequence databases using mass spectrometry data.** Electrophoresis, 20 (18) (1999), pp. 3551-3567

[25] M. Vaudel, H. Barsnes, F.S. Berven, A. Sickmann, L. Martens. **SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches.** Proteomics, 11 (5) (2011), pp. 996-999

[26] D. Fenyö, R.C. Beavis. **A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes.** Anal. Chem., 75 (4) (2003), pp. 768-774

[27] S. Kim, P.A. Pevzner. **MS-GF + makes progress towards a universal database search tool for proteomics.** Nat. Commun., 5 (2014), p. 5277

[28] V. Dorfer, P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler, K. Mechtler. **MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra.** J. Proteome Res., 13 (8) (2014), pp. 3679-3684

[29] D.L. Tabb, C.G. Fernando, M.C. Chambers. **MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis.** J. Proteome Res., 6 (2) (2007), pp. 654-661

[30] J.K. Eng, T.A. Jahan, M.R. Hoopmann. **Comet: an open-source MS/MS sequence database search tool.** Proteomics, 13 (1) (2013), pp. 22-24

[31] B.J. Diament, W.S. Noble. **Faster SEQUEST searching for peptide identification from tandem mass spectra.** J. Proteome Res., 10 (9) (2011), pp. 3871-3879

[32] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann. **Andromeda: a peptide search engine integrated into the MaxQuant environment.** J. Proteome Res., 10 (4) (2011), pp. 1794-1805

[33] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant. **Open mass spectrometry search algorithm.** J. Proteome Res., 3 (5) (2004), pp. 958-964

[34] B. Ma. **Novor: real-time peptide de novo sequencing software.** J. Am. Soc. Mass Spectrom., 26 (11) (2015), pp. 1885-1894

[35] D.L. Tabb, Z.-Q. Ma, D.B. Martin, A.-J.L. Ham, M.C. Chambers. **DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring.** J. Proteome Res., 7 (9) (2008), pp. 3838-3846

[36] M. Vaudel, J.M. Burkhart, R.P. Zahedi, E. Oveland, F.S. Berven, A. Sickmann, L. Martens, H. Barsnes. **PeptideShaker enables reanalysis of MS-derived proteomics data sets.** Nat. Biotechnol., 33 (1) (2015), pp. 22-24

[37] J. Cox, M. Mann. **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.** Nat. Biotechnol., 26 (12) (2008), pp. 1367-1372

[38] R. Aebersold, M. Mann. **Mass spectrometry-based proteomics.** Nature, 422 (6928) (2003), pp. 198-207

[39] B. Cañas, C. Piñeiro, E. Calvo, D. López-Ferrer, J.M. Gallardo. **Trends in sample preparation for classical and second generation proteomics.** J. Chromatogr. A, 1153 (1–2) (2007), pp. 235-258

[40] J. Rodriguez, N. Gupta, R.D. Smith, P.A. Pevzner. **Does trypsin cut before proline?**. J. Proteome Res., 7 (1) (2008), pp. 300-305

[41] D.A. Abaye, F.S. Pullen, B.V. Nielsen. **Peptide polarity and the position of arginine as sources of selectivity during positive electrospray ionisation mass spectrometry.** Rapid Commun. Mass Spectrom., 25 (23) (2011), pp. 3597-3608

[42] R. King, R. Bonfiglio, C. Fernandez-Metzler, C. Miller-Stein, T. Olah. **Mechanistic investigation of ionization suppression in electrospray ionization.** J. Am. Soc. Mass Spectrom., 11 (11) (2000), pp. 942-950

[43] A. Hirabayashi, M. Ishimaru, N. Manri, T. Yokosuka, H. Hanzawa. **Detection of potential ion suppression for peptide analysis in nanoflow liquid chromatography/mass spectrometry.** Rapid Commun. Mass Spectrom., 21 (17) (2007), pp. 2860-2866

[44] P. Schliekelman, S. Liu. **Quantifying the effect of competition for detection between coeluting peptides on detection probabilities in mass-spectrometry-based proteomics.** J. Proteome Res., 13 (2) (2013), pp. 348-361

[45] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N.G. Ahn, W.M. Old. **Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies.** J. Proteome Res., 9 (8) (2010), pp. 4152-4160

[46] A. Michalski, J. Cox, M. Mann. **More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC − MS/MS.** J. Proteome Res., 10 (4) (2011), pp. 1785-1793

[47] V. Gorshkov, S.Y.K. Hotta, T. Verano-Braga, F. Kjeldsen. **Peptide de novo sequencing of mixture tandem mass spectra.** Proteomics, 16 (18) (2016), pp. 2470-2479

[48] M. Blein-Nicolas, M. Zivy. **Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics.** Biochim. Biophys. Acta, Proteins Proteomics, 1864 (8) (2016), pp. 883-895

[49] L. Dicker, X. Lin, A.R. Ivanov. **Increased power for the analysis of label-free LC-MS/MS proteomics data by combining spectral counts and peptide peak attributes.** Mol. Cell. Proteomics, 9 (12) (2010), pp. 2704-2718

[50] H. Liu, R.G. Sadygov, J.R. Yates. **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** Anal. Chem., 76 (14) (2004), pp. 4193-4201

[51] W.M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K.G. Pierce, A. Mendoza, J.R. Sevinsky, K.A. Resing, N.G. Ahn. **Comparison of label-free methods for quantifying human proteins by shotgun proteomics.** Mol. Cell. Proteomics, 4 (10) (2005), pp. 1487-1502

[52] Y. Zhang, Z. Wen, M.P. Washburn, L. Florens. **Effect of dynamic exclusion duration on spectral count based quantitative proteomics.** Anal. Chem., 81 (15) (2009), pp. 6317-6326

[53] M. Bantscheff, S. Lemeer, M. Savitski, B. Kuster. **Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present.** Anal. Bioanal. Chem., 404 (4) (2012), pp. 939-965

[54] T.I. Milac, T.W. Randolph, P. Wang. **Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies, statistics and its interface.** 5(1) (2012), pp. 75-87

[55] J.F. Krey, P.A. Wilmarth, J.-B. Shin, J. Klimek, N.E. Sherman, E.D. Jeffery, D. Choi, L.L. David, P.G. Barr-Gillespie. **Accurate label-free protein quantitation with high- and low-resolution mass spectrometers.** J. Proteome Res., 13 (2) (2014), pp. 1034-1044

[56] Y. Zhang, B.R. Fonslow, B. Shan, M.-C. Baek, J.R. Yates. **Protein analysis by shotgun/bottom-up proteomics.** Chem. Rev., 113 (4) (2013), pp. 2343-2394

[57] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, O. Vitek. **Protein quantification in label-free LC-MS experiments.** J. Proteome Res., 8 (11) (2009), pp. 5275-5284

[58] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach. **Global quantification of mammalian gene expression control.** Nature, 473 (7347) (2011), pp. 337-342

[59] Y.-Y. Chen, M.C. Chambers, M. Li, A.-J.L. Ham, J.L. Turner, B. Zhang, D.L. Tabb. **IDPQuantify: combining precursor intensity with spectral counts for protein and peptide quantification.** J. Proteome Res., 12 (9) (2013), pp. 4111-4121

[60] R.E. Higgs, M.D. Knierman, V. Gelfanova, J.P. Butler, J.E. Hale. **Comprehensive label-free method for the relative quantification of proteins from biological samples.** J. Proteome Res., 4 (4) (2005), pp. 1442-1450

[61] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M.A. Gillette, S.A. Carr. **PEPPeR, a platform for experimental proteomic pattern recognition.** Mol. Cell. Proteomics, 5 (10) (2006), pp. 1927-1941

[62] J. Malmstrom, M. Beck, A. Schmidt, V. Lange, E.W. Deutsch, R. Aebersold. **Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans.*** Nature, 460 (7256) (2009), pp. 762-765

[63] B. Dost, N. Bandeira, X. Li, Z. Shen, S.P. Briggs, V. Bafna. **Accurate mass spectrometry based protein quantification via shared peptides.** J. Comput. Biol., 19 (4) (2012), pp. 337-348

[64] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann. **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.** Mol. Cell. Proteomics, 13 (9) (2014), pp. 2513-2526

[65] D.S. Daly, K.K. Anderson, E.A. Panisko, S.O. Purvine, R. Fang, M.E. Monroe, S.E. Baker. **Mixed-effects statistical model for comparative LC − MS proteomics studies.** J. Proteome Res., 7 (3) (2008), pp. 1209-1217

[66] M. Choi, C.-Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, O. Vitek. **MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.** Bioinformatics, 30 (17) (2014), pp. 2524-2526

[67] Y.V. Bukhman, M. Dharsee, R. Ewing, P. Chu, T. Topaloglou, T. Le Bihan, T. Goh, H. Duewel, I.I. Stewart, J.R. Wisniewski, N.F. Ng. **Design and analysis of quantitative differential proteomics investigations using LC-MS technology.** J. Bioinforma. Comput. Biol., 6 (1) (2008), pp. 107-123

[68] M. Blein-Nicolas, H. Xu, D. de Vienne, C. Giraud, S. Huet, M. Zivy. **Including shared peptides for estimating protein abundances: a significant improvement for quantitative proteomics.** Proteomics, 12 (18) (2012), pp. 2797-2801

[69] S. Gerster, T. Kwon, C. Ludwig, M. Matondo, C. Vogel, E.M. Marcotte, R. Aebersold, P. Bühlmann. **Statistical approach to protein quantification.** Mol. Cell. Proteomics, 13 (2) (2014), pp. 666-677

[70] S.Y. Ryu, W.-J. Qian, D.G. Camp, R.D. Smith, R.G. Tompkins, R.W. Davis, W. Xiao. **Detecting differential protein expression in large-scale population proteomics.** Bioinformatics, 30 (19) (2014), pp. 2741-2746

[71] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J.N. Adkins, C. Ansong, F. Heffron, T.O. Metz, W.-J. Qian, H. Yoon, R.D. Smith, A.R. Dabney. **A statistical framework for protein quantitation in bottom-up MS-based proteomics.** Bioinformatics, 25 (16) (2009), pp. 2028-2034

[72] T. Taverner, Y.V. Karpievitch, A.D. Polpitiya, J.N. Brown, A.R. Dabney, G.A. Anderson, R.D. Smith. **DanteR: an extensible R-based tool for quantitative analysis of -omics data.** Bioinformatics, 28 (18) (2012), pp. 2404-2406

[73] L.J.E. Goeminne, A. Argentini, L. Martens, L. Clement. **Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines.** J. Proteome Res., 14 (6) (2015), pp. 2457-2465

[74] L.J.E. Goeminne, K. Gevaert, L. Clement. **Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics.** Mol. Cell. Proteomics, 15 (2) (2016), pp. 657-668

[75] P. Blainey, M. Krzywinski, N. Altman. **Points of significance: replication.** Nat. Methods, 11 (9) (2014), pp. 879-880

[76] D.L. Vaux, F. Fidler, G. Cumming. **Replicates and repeats—what is the difference and is it significant?: a brief discussion of statistics and experimental design.** EMBO Rep., 13 (4) (2012), pp. 291-296

[77] M. Krzywinski, N. Altman. **Points of significance: analysis of variance and blocking.** Nat. Methods, 11 (7) (2014), pp. 699-700

[78] O. Serang, W. Noble. **A Review of Statistical Methods for Protein Identification Using Tandem Mass Spectrometry, Statistics and its Interface.** 5(1) (2012), pp. 3-20

[79] S.J. Callister, R.C. Barry, J.N. Adkins, E.T. Johnson, W.-j. Qian, B.-J.M. Webb-Robertson, R.D. Smith, M.S. Lipton. **Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics.** J. Proteome Res., 5 (2) (2006), pp. 277-286

[80] K. Hodge, S.T. Have, L. Hutton, A.I. Lamond. **Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS.** J. Proteome, 88 (2013), pp. 92-103

[81] N. Altman, M. Krzywinski. **Points of significance: simple linear regression.** Nat. Methods, 12 (11) (2015), pp. 999-1000

[82] N. Altman, M. Krzywinski. **Points of significance: sources of variation.** Nat. Methods, 12 (1) (2015), pp. 5-6

[83] W.S. Noble. **How does multiple testing correction work?.** Nat. Biotechnol., 27 (12) (2009), pp. 1135-1137

[84] A.P. Diz, A. Carvajal-Rodríguez, D.O.F. Skibinski. **Multiple hypothesis testing in proteomics: a strategy for experimental work.** Mol. Cell. Proteomics, 10 (3) (2011)

[85] Y. Benjamini, Y. Hochberg. **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** J. R. Stat. Soc. Ser. B Methodol., 57 (1) (1995), pp. 289-300

[86] R Core Team. **R: a language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria (2016)

[87] RStudio Team. **RStudio: Integrated Development for R.** RStudio, Inc., Boston, MA (2015)

[88] E. Ramond, G. Gesbert, I.C. Guerrera, C. Chhuon, M. Dupuis, M. Rigard, T. Henry, M. Barel, A. Charbit. **Importance of host cell arginine uptake in *Francisella* phagosomal escape and ribosomal protein amounts.** Mol. Cell. Proteomics, 14 (4) (2015), pp. 870-881

[89] A.G. Paulovich, D. Billheimer, A.-J.L. Ham, L. Vega-Montoto, P.A. Rudnick, D.L. Tabb, P. Wang, R.K. Blackman, D.M. Bunk, H.L. Cardasis, K.R. Clauser, C.R. Kinsinger, B. Schilling, T.J. Tegeler, A.M. Variyath, M. Wang, J.R. Whiteaker, L.J. Zimmerman, D. Fenyo, S.A. Carr, S.J. Fisher, B.W. Gibson, M. Mesri, T.A. Neubert, F.E. Regnier, H. Rodriguez, C. Spiegelman, S.E. Stein, P. Tempst, D.C. Liebler. **Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance.** Mol. Cell. Proteomics, 9 (2) (2010), pp. 242-254

[90] D. Amaratunga, J. Cabrera. **Analysis of data from viral DNA microchips.** J. Am. Stat. Assoc., 96 (456) (2001), pp. 1161-1170

[91] L. Gatto, K.S. Lilley. **MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.** Bioinformatics, 28 (2) (2012), pp. 288-289

[92] N. Gupta, P.A. Pevzner. **False discovery rates of protein identifications: a strike against the two-peptide rule.** J. Proteome Res., 8 (9) (2009), pp. 4173-4181

[93] M. Krzywinski, N. Altman. **Points of significance: visualizing samples with box plots.** Nat. Methods, 11 (2) (2014), pp. 119-120

[94] C. Lazar, L. Gatto, M. Ferro, C. Bruley, T. Burger. **Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies.** J. Proteome Res., 15 (4) (2016), pp. 1116-1125

[95] B.-J.M. Webb-Robertson, H.K. Wiberg, M.M. Matzke, J.N. Brown, J. Wang, J.E. McDermott, R.D. Smith, K.D. Rodland, T.O. Metz, J.G. Pounds, K.M. Waters. **Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics.** J. Proteome Res., 14 (5) (2015), pp. 1993-2001

[96] A. Bilbao, E. Varesio, J. Luban, C. Strambio-De-Castillia, G. Hopfgartner, M. Müller, F. Lisacek. **Processing strategies and software solutions for data-independent acquisition in mass spectrometry.** Proteomics, 15 (5–6) (2015), pp. 964-980

[97] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, J.A. Vizcaíno. **Making proteomics data accessible and reusable: current state of proteomics databases and repositories.** Proteomics, 15 (5–6) (2015), pp. 930-950

[98] M. Larance, A.I. Lamond. **Multidimensional proteomics for cell biology.** Nat. Rev. Mol. Cell Biol., 16 (5) (2015), pp. 269-280

[99] R. Aebersold, M. Mann. **Mass-spectrometric exploration of proteome structure and function.** Nature, 537 (7620) (2016), pp. 347-355

[100] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W.H. Tang, A. Römpp, S. Neumann, A.D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, E.W. Deutsch. **mzML—a community standard for mass spectrometry data.** Mol. Cell. Proteomics, 10 (1) (2011)

[101] M. Walzer, D. Qi, G. Mayer, J. Uszkoreit, M. Eisenacher, T. Sachsenberg, F.F. Gonzalez-Galarza, J. Fan, C. Bessant, E.W. Deutsch, F. Reisinger, J.A. Vizcaíno, J.A. Medina-Aunon, J.P. Albar, O. Kohlbacher, A.R. Jones. **The mzQuantML data standard for mass spectrometry–based quantitative studies in proteomics.** Mol. Cell. Proteomics, 12 (8) (2013), pp. 2332-2340

[102] A. Hu, W.S. Noble, A. Wolf-Yadlin

**Technical advances in proteomics: new developments in data-independent acquisition, F1000Research 5.** (2016) (F1000 Faculty Rev-419)

**Ludger J.E. Goeminne** is a PhD student working on differential proteomics in the StatOmics Lab headed by Prof. Lieven Clement and the proteomics lab headed by Prof. Kris Gevaert. In 2015, he demonstrated in the Journal of Proteome Research that peptide-based models outperform summarization-based approaches. He further developed an optimized peptide-based modeling approach, published in Molecular and Cellular Proteomics (2016). Ludger also co-authored the 2016 Nature Methods paper by Argentini et al. Additionally, Ludger won the best flash talk presentation award at Eubic Winter School 2017 and he is a member of the vibes 2017 PhD Symposium Organizing Committee (www.vibes2017.com).

**Kris Gevaert** holds a PhD in Biotechnology (2000) at Ghent University (Belgium) and is currently a Full Professor at Ghent University and associate department director of the VIB-UGent Center for Medical Biotechnology (http://mbc.vib-ugent.be/). His group published more than 280 papers and several book chapters on the development and applications of proteomics techniques in several areas of biomedical and life sciences research.

**Lieven Clement** is Assistant Professor of Statistical Genomics at Ghent University. His research group focuses on developing statistical methods for omics profiling (proteomics, transcriptomics and (meta)genomics) with mass spectrometry, next-generation sequencing (NGS) and digital PCR platforms. The statistical tools and software are motivated by practical applications in biology, biotechnology and biomedical research and build upon a strong collaboration with research groups in the life sciences. He also leverages his expertise to translational research and serves as an expert in genomics projects of the Belgian Health Care Knowledge Center (KCE).

## Appendix (not a part of the original article, but part of my PhD thesis)

We propose the same peptide-based regression model as in Goeminne *et al.* (2016) [1]:

$$y_{pr} = \boldsymbol{x}_{pr}\boldsymbol{\beta} + \beta_p^{\text{peptide}} + u_r^{\text{run}} + \varepsilon_{pr}$$

Herein, $\boldsymbol{x}_{pr}$ is a row matrix with the covariate pattern related to peptide $p$ in run $r$, $\boldsymbol{\beta} = \left[\beta^0, \beta_1^1 \dots, \beta_{m_1}^1 \dots, \beta_{M_1}^1, \dots, \beta_{m_g}^g, \dots, \beta_{M_g}^g, \dots, \beta_{M_G}^G\right]^{\mathrm{T}}$ is a vector with $1 + M = 1 + \sum_{g=1}^{G} M_g$ parameters denoting the effects of $M$ predictors corresponding to $G$ covariates. $\beta_p^{\text{peptide}}$ is a peptide-specific effect for peptide $p$, $u_r^{\text{run}}$ a random run effect to account for within-run correlation, with $u_r^{\text{run}} \sim \mathrm{N}(0, \sigma_u^2)$. $\varepsilon_{pr} \sim \mathrm{N}(0, \sigma^2)$ is a random error term.

The new version of MSqRob allows the user to put the same ridge penalty on multiple covariates because traditional ridge regression has a single ridge penalty $\lambda$ that is equal for all fixed effects instead of separate ridge penalties for each covariate. We thus assume all ridge parameters to originate from the same distribution: $\beta_{m_g}^g \sim \mathrm{N}(0, \sigma^2/\lambda)$ for $m = 1, \dots, M_g$ and $g = 1, \dots, G$. Note that we still require a separate ridge penalty for the peptide effects $\beta_p^{\text{peptide}} \sim \mathrm{N}(0, \sigma^2/\lambda_{\text{peptide}})$ because of their large effect sizes.

When naively imposing a single ridge penalty on multiple covariates, the amount of shrinkage is influenced by the scale of the predictors and the model's parameterization (e.g. the choice of the reference class can impact on the ridge penalty). To ensure that the size of the ridge penalty is independent of the model's parameterization, we perform a QR-decomposition on the part of the design matrix that corresponds to the fixed effect covariates, $\boldsymbol{X}^{\text{fixed}}$.

$$\boldsymbol{X}^{\text{fixed}} = \boldsymbol{QR}$$

Herein, the $\boldsymbol{Q}$-matrix is an orthogonal matrix and can be used as a rescaled version of the original design matrix $\boldsymbol{X}^{\text{fixed}}$. The $\boldsymbol{R}$-matrix is an upper triangular matrix.

Subsequently, $\boldsymbol{X}^{\text{fixed}}$ is replaced by the $\boldsymbol{Q}$-matrix prior to the lme4 mixed model fitting. During statistical inference, the part of the design matrix corresponding to the fixed effects is post-multiplied with the $\boldsymbol{R}$-matrix to return to the original scale.

## Implementation

The lme4 R package does not allow to impose a single ridge penalty over a group of multiple covariates. Therefore, we set up an lme4 model with a mock random effect that has the same number of levels as there are levels for the fixed effects that are shrunken together:

```
parsedFormula <- lFormula(y~1+(1|ridgeGroup)+…)
```

Then, we construct the part of the design matrix that corresponds to the shrunken fixed effects and perform QR-decomposition

$X_{ridgeGroup}$ <- model.matrix(…)

$Q_{ridgeGroup}$ <- qr.Q(qr($X_{ridgeGroup}$))

$R_{ridgeGroup}$ <- qr.R(qr($X_{ridgeGroup}$))

Next, we change the part of the design matrix in the parsedFormula that corresponds to the shrunken fixed effects:

```
parsedFormula$reTrms          <-          within(parsedFormula$reTrms,
{Zt[ridgeGroupindices,] <- t(Q_ridgeGroup)})
```

And finally fit the model:

```
devianceFunction <- do.call(mkLmerDevfun, parsedFormula)

optimizerOutput <- optimizeLmer(devianceFunction)

mRidge <- mkMerMod(

                   rho = environment(devianceFunction),

                   opt = optimizerOutput,

                   reTrms = parsedFormula$reTrms,

                   fr = parsedFormula$fr)
```

When doing inference, the part of the design matrix that was replaced with $Q_{ridgeGroup}$, is post-multiplied with $R_{ridgeGroup}$. A similar procedure to create a single ridge penalty for multiple covariates in lm4 models has also exploited by the gamm4 R package [2].

## References for the Appendix

1. Goeminne, L.J.E., K. Gevaert, and L. Clement, *Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics.* Molecular & Cellular Proteomics, 2016. **15**(2): p. 657-668.
2. Wood, S.N., *Generalized Additive Models: An Introduction with R*. 2006: Chapman and Hall/CRC