UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

State-of-the-art Mass Spectrometry quantification data normalization using ANOVA Models

Piotr Prostko

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Dirk VALKENBORG

MENTOR :

Mevrouw Annelies AGTEN

De heer Joris VAN HOUTVEN

2019
2020

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

State-of-the-art Mass Spectrometry quantification data normalization using ANOVA Models

Piotr Prostko

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Dirk VALKENBORG

MENTOR :

Mevrouw Annelies AGTEN

De heer Joris VAN HOUTVEN

List of Abbreviations

AA	Amino acid
ANOVA	ANalysis Of VAriance
B-H	Benjamini-Hochberg
BM	Bone marrow cells to produce macrophages
FC	Fold change
FDR	False discovery rate
ICAT	Isotope-Coded Affinity Tag
iTRAQ	Isobaric Tags for Relative and Absolute Quantitation
LC	Liquid chromatography
LMM	Linear Mixed Model
ML	Maximum Likelihood
MS	Mass spectrometry
MS^2	Tandem mass spectrometry
m/z	Mass-to-charge ratio
NGS	Next Generation Sequencing
PCA	Principal components analysis
PFF	Peptide Fragment Fingerprinting
PM	Peritoneal macrophages
PMF	Peptide Mass Fingerprinting
PSM	Peptide spectrum match
PTM	Post translational modification
REML	REstricted Maximum Likelihood
RT	Retention time
TAM	Tumor-associated macrophages
TMT	Tandem Mass Tags

Abstract

Isobaric mass labeling is an attractive strategy in MS-based quantitative proteomics offering peptide identification and quantification in multiple biological samples within a single machine run. This feature allows reducing technical variability, but the normalization of recorded reporter ion intensities still remains to be an important step of any analysis workflow.

Considering the vast amounts of proteomic data, fast and data-driven normalization methods are becoming more popular in analytical workflows applied in high-throughput laboratories. On the other hand, more time consuming but also more statistically thorough, and as such potentially more accurate methods like linear models can be employed to normalize peptide abundance data. Therefore, the goal of this research is to compare a data-driven technique CONSTANd with one of the state-of-the-art data normalization methods based on ANOVA. The methods have been evaluated on two real-life datasets: a dataset concerning samples consisting of three different cell types taken from mice with orthotopic non-small cell lung adenocarcinomas and analyzed with three MS^2 runs (true proteins abundance unknown), and a single MS^2 run spiked-in experiment. We compared the two methods in terms of normalization capabilities (via boxplot, MA plot, PCA plot and heatmap), computation time and their impact on differential expression analysis results ($\log_2$ fold change and p-value). The obtained results do not support the claim that the state-of-the-art ANOVA based method outperforms CONSTANd. However, further research based on simulated data or a spiked-in dataset with a large number of samples is needed to understand the influence of the normalization procedures on the false positive and false negative proportions in a differential expression analysis.

Contents

1 Introduction

Today we are witnessing continuous progress in medicine and health care. Every year novel, more safe and more efficacious treatments are introduced to clinics around the world, or new indications are being found for drugs already on the market. Another novelty is that nowadays there is more focus on finding biomarkers, i.e. molecules allowing the detection of a disease of interest, prognosis of its evolution over time or prediction of a potential response to the treatment for a particular disease. However, all those important changes in patient management would not be possible (or at least not possible on a large scale) without advances in analytical techniques used in studying underlying molecular and cellular processes taking place in diseased subjects.

Before the development of high-throughput analytical techniques, wet-lab studies aiming at finding promising active molecules relied mostly on the time consuming process of manual curation of biological samples by a skilled analyst. Such approach could focus only on a very limited number of molecular entities studied simultaneously and would most likely miss complex dependencies between numerous molecules. Instead, modern studies make use of techniques like DNA microarrays, next generation sequencing (NGS) or mass spectrometry (MS). Such cutting-edge machinery generate vast amounts of data, commonly described with the suffix "-omics". Omics analyses offer a system-wide understanding of a biological problem at hand, but this comes at a price of expensive hardware and the requirement of using sophisticated data processing and statistical algorithms to account for various sources of confounding (due to physical and chemical processes operating inside those complex machines). Here in this report, MS-based proteomics will be discussed.

Proteins are the end product of the process named "the central dogma of molecular biology", which involves turning genetic information coded in DNA into molecules ready to perform specific functions in a cell. Moreover, proteins are composed of smaller building blocks called peptides, and a peptide is a chain of amino acids (AA).
Proteins can be further characterized based on their primary (AA sequence), secondary (folding of the AA sequence and creation of local structures), tertiary (the overall shape of a single peptide chain and the formation of secondary structures into local domains) and quaternary structure (a merge of several polypeptide chains resulting in the target protein). This shows that a protein is the result of complex, multilevel processes, throughout which modifications may occur in several places, thus making identification and quantification (of its expression) a challenging task.

For the large scale identification and quantification of proteins in specific cells, tissues or organisms experiencing some interesting physiological and environmental conditions, mass spectrometry is often used. A mass spectrometer is an expensive scale which is able to weigh very light objects like atoms or molecules and measure their abundance in a sample by separating them according to mass-to-charge ratio (m/z).
Further, a typical spectrometer is build from three main components: ion source, mass analyzer, and ion detector. Several variants of each of these parts exist, leading to a different MS procedure with its own technical nuances. As an illustration of mechanisms operating in a spectrometer, consider the following procedure.

First, a sample with molecules is inserted into a ion source, where molecules are hit with electrons and become positively charged. Then, in the mass analyzer part, those ions are exposed to an electrical field and accelerated for a short while. Afterwards those particles enter the field-free vacuum tube and travel freely through it until they reach (hit) the

detector. The time of this travel depends on the mass-to-charge ratio - assuming the same charge state, heavier particles should arrive later than particles with a lower mass. Finally, ions are recorded by the detector and the relative intensity at a particular m/z region is outputted. A plot showing relative intensities versus m/z is called a (mass) spectrum (see Figure 1).

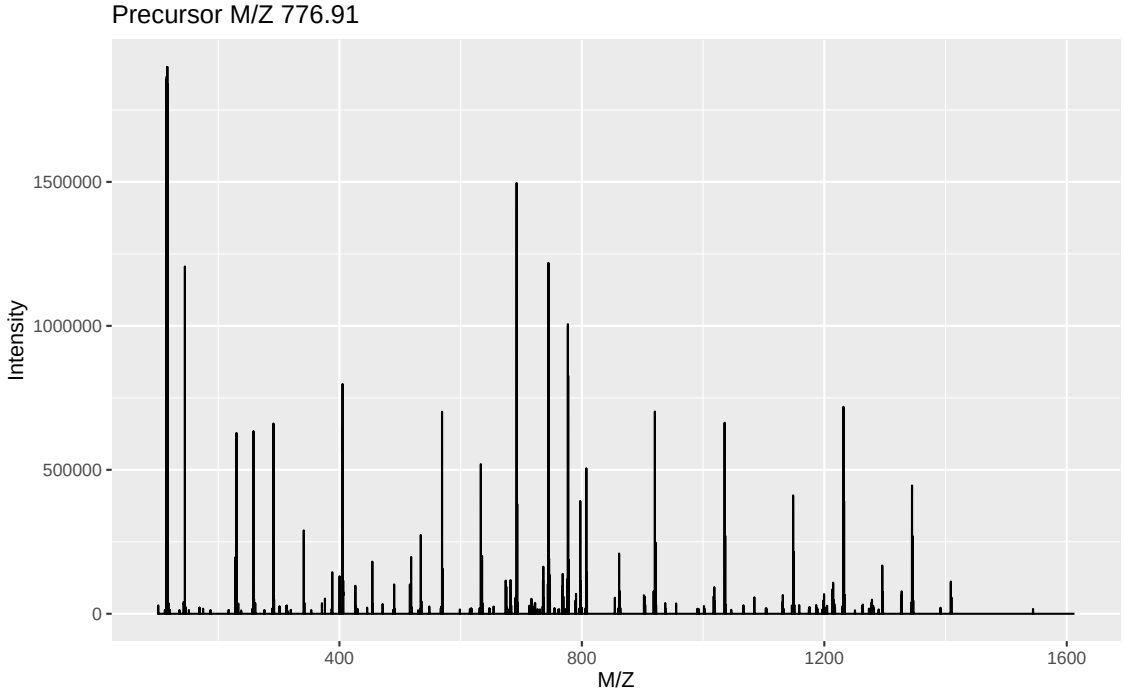

Figure 1: An example of a spectrum obtained from a MS experiment.

Having this simple intuition on how a mass spectrometer works, basic steps in a typical MS analysis workflow will be briefly discussed (Figure 2).

Proteins are large biomolecules, as such, it is not straightforward to analyze intact proteins with current technology. Therefore the first step in the workflow is enzymatic digestion, for instance using trypsin, which results in a complex mixture of peptides. In the next step, the mixture is loaded into the column of liquid chromatography (LC) in order to reduce its complexity. Indeed, LC allows peptide separation based on different physico-chemical properties of the molecules in the sample. Simplistically, it means that the solution containing different peptides elutes from the LC column at different times (retention time, RT), so that a more homogenous peptide mixture can be analyzed with MS.

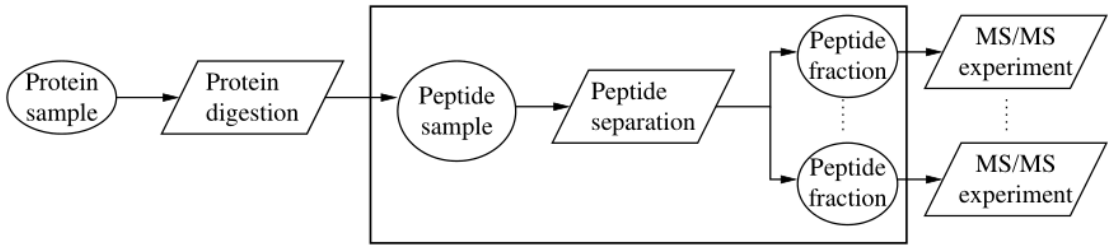

Figure 2: Scheme of main steps in a MS experiment. Source: after Eidhammer et al., 2007, chapter 4. [1]

Now, two general approaches for MS based peptide characterization can be distinguished: Peptide Mass Fingerprinting (PMF) or Peptide Fragment Fingerprinting (PFF). The former focuses on the identification based on intact peptide masses, but it has some drawbacks and limitations e.g. it cannot handle peptide modifications or it cannot handle well peptides with nearly identical mass. PFF involves one extra step - breaking peptides into smaller bits (fragments) and analyzing them with one additional MS run, which explains its name - tandem mass spectrometry (MS^2). MS^2 allows for more accurate peptide identifications as it yields mass information on a more detailed level, namely, ion fragment level.

The sample content, in terms of proteins and peptides, is a priori unknown, therefore a collection of spectra (either obtained from MS or MS^2) needs to be processed further. To shed some more light on this, assume that a MS^2 experiment has been carried out. To begin this process, the observed masses from the first MS scan (parent scan) are compared with peptide masses included in a protein sequence database (for example Swiss-Prot or TrEMBL). Candidate peptides outside the user specified mass tolerance are filtered out. For the remaining candidate peptides, theoretical spectra can be computed. Peptide by peptide, this is done by considering all possible ion fragments, charge states up to a specified threshold. Next, the agreement between peaks in the theoretical spectra of the candidate peptides and peaks in the acquired spectra is evaluated with a preferred scoring algorithm, for example Mascot. Such an algorithm yields a numerical score that assess the quality of the peptide spectrum match (PSM). Then, usually only one candidate peptide per spectrum, the one with the highest score, is selected for the downstream analysis. However, the best match could be true or false identification, therefore one should also control the rate of false identifications in the entire data set. This goal can be achieved with, for instance, target-decoy database searching approach (Elias et al., [2]). Finally, the result of this briefly described procedure is a list of peptides identified in the experiment(s).

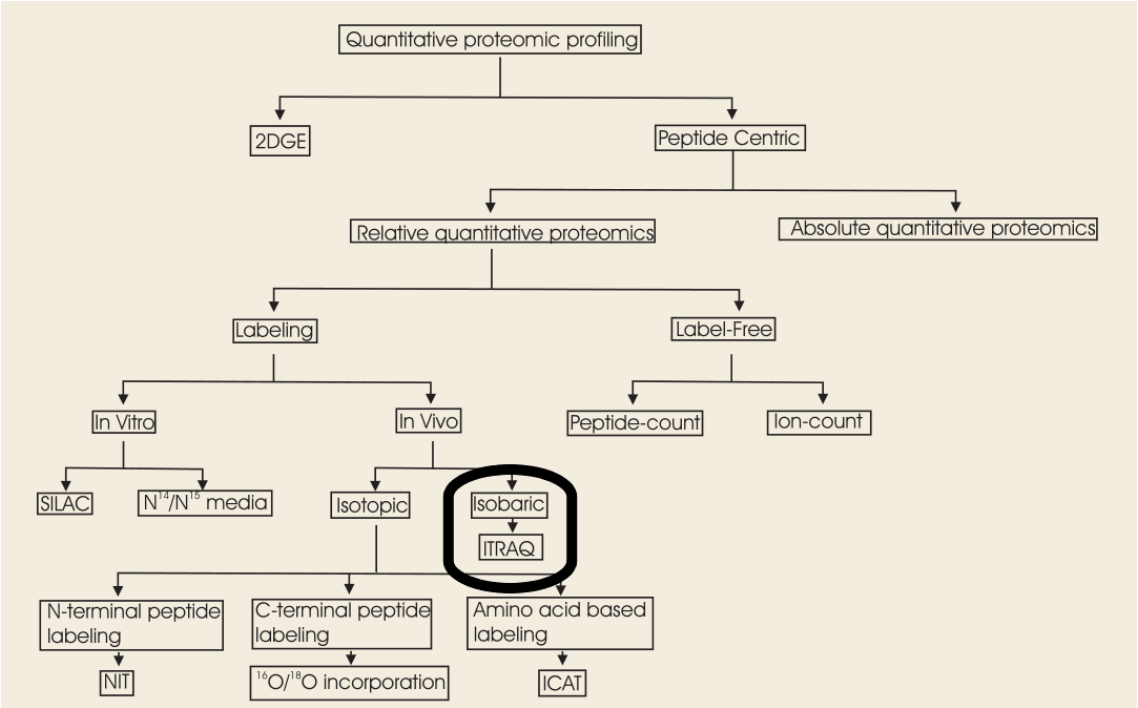

Figure 3: An overview of methods in quantitative MS based proteomics. The focus of this research is isobaric labeling. Source: after "Analysis of Protein Expression" course at UHasselt.

Besides the identification task, interest often lies in a relative quantitation of the analyzed sample, for instance, one may want to find proteins with significantly different relative abundances between two (or more) treatment arms. As shown in Figure 3, there are different ways to achieve this goal: label-free methods and various labeling strategies such as ICAT (Isotope-Coded Affinity Tag); O^{16}/O^{18} (heavy oxygen water); ITRAQ (Isobaric Tags for Relative and Absolute Quantitation) or TMT (Tandem Mass Tags). iTRAQ and TMT are both examples of isobaric mass labels, which allow simultaneous identification and quantification of peptides from multiple biological samples within a single MS^2 run (iTRAQ 4- and 8-plex or 2-,6-,10-,11-plex available in case of TMT).

In this approach, just after the enzymatic digestion of proteins and other sample preparation procedures, samples from different objects/conditions are treated with isobaric labels, then mixed and combined together into one sample, which is further analyzed with a single MS^2 run. As depicted in Figure 4A, each isobaric label consists of three parts: reporter region, balance region and peptide reactive group. The peptide reactive group is responsible for attaching the labels to the peptides, while the role of the reporter region is to distinguish the combined samples from each other, as each of them has a different mass (e.g. 114 Da, 115 Da, 116 Da, 117 Da, in case of ITRAQ 4-plex). Lastly, the balance region ensures equal mass of each label in order not to introduce a mass shift in the collected MS spectra (hence the adjective "isobaric"). In the fragmentation step of MS^2 the reporter region is cleaved off and analyzed as other fragment ions. The recorded intensities of reporter ions reflect the abundance of each labeled sample (Figure 4B).

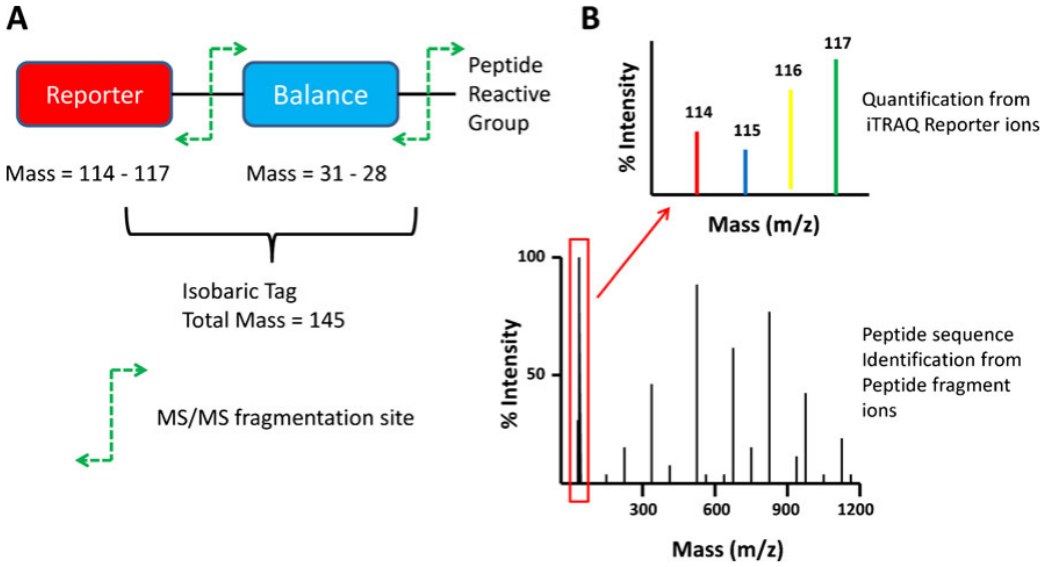

Figure 4: A detailed view on iTRAQ labeling. Source: after Xu et al. [3].

The result, after some additional data processing, of an iTRAQ or TMT experiment is a quantification matrix in which rows correspond to the identified peptides and columns represent the analyzed samples. However, this data is affected by systematic errors due to, for example, sample handling, pipetting errors, machine settings or biological effects. Therefore, there is an eminent need to normalize the data by removing these unwanted technical effects, while retaining only biologically relevant effects. Correct data normalization is one of the key steps to obtain meaningful results.

1.1 Proteomic data structure

Prior to discussing CONSTANd and the ANOVA based approach, it is worthwhile to explain the structure of data resulting from the process of peptide identification in an MS^2 experiment. This will allow the reader to understand what data features/structures can be taken into account by a given analysis approach.

As indicated previously, in an isobaric labeling experiment, multiple samples can be studied within a single machine run. If the number of samples to investigate is larger than the size of the TMT/iTRAQ kit, multiple runs of the mass spectrometer are needed. Therefore the first layer of the data structure corresponds to the study design. Figure 8 displays an example of a study design with three MS^2 runs in a TMT 6-plex experiment.

Consider now the content of only one of the study samples. As displayed in Figure 5, there is a multilevel hierarchy in this type of data. Starting from the top, thousands of proteins can be detected within the sample. Those protein sequences consist of multiple peptides. Note that some peptides can occur in more than one protein, and such peptides are called "shared peptides". In order not to introduce additional complexity, shared peptides were discarded from differential expression analysis (DEA). Usually, investigators formulate their research questions in terms of the significant changes in protein expression levels between two or more groups of subjects. Hence, within a sample, peptides can be perceived as repeated measurements of proteins. One level down, post translational modifications (PTM), giving rise to different variants of peptides, may occur during the protein synthesis. Likewise, a certain peptide may have various charge states and finally, the aforementioned peptide variants may elute from the LC column at multiple points in time, called retention times (RT). As pointed out by the green bracket in Figure 5, we

can look at PTMs, charge state and RT as repeated measurements on the peptide level. It is often assumed that these repeated measurements (at least those due to charge state and RT) are rather technical or artificial replicates of the peptide and should not add much of additional information. Therefore, in many analytical workflows for quantitative proteomics, there is an aggregation step to remove redundancy due to RT, charge states and PTMs, resulting in only one quantification values per peptide within the sample.

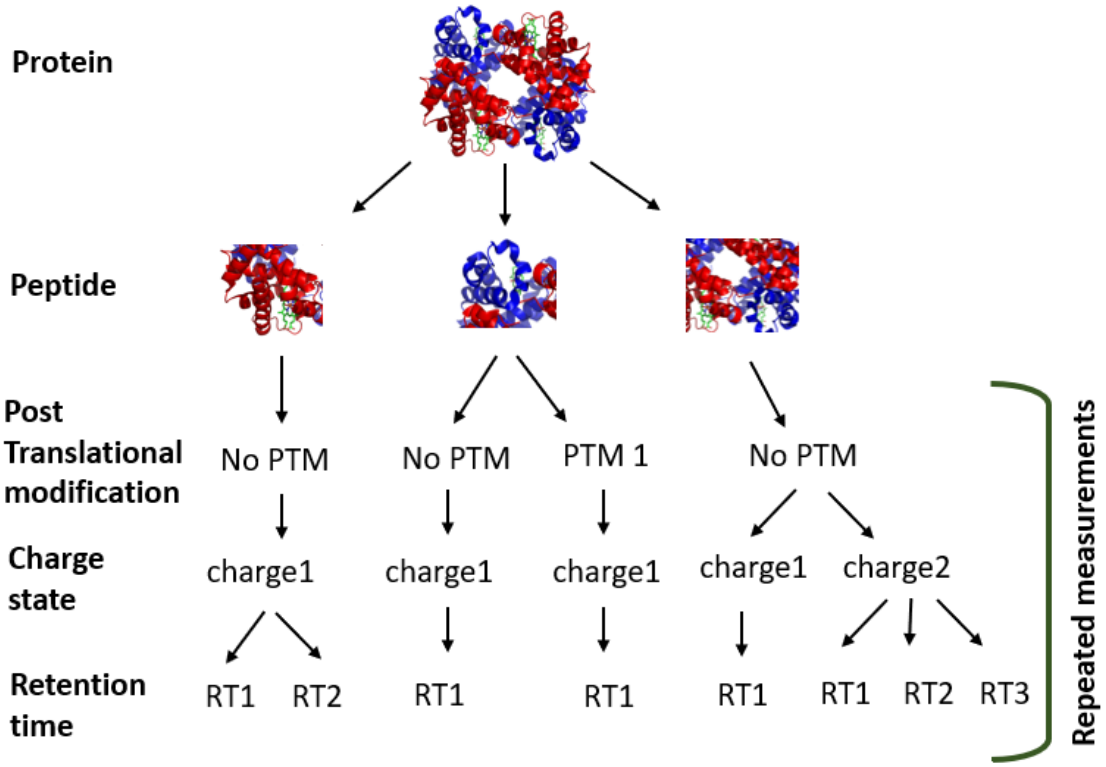

Figure 5: A visual description of the proteomic data hierarchy. Within a sample, one protein can be represented by multiple peptides, and those might be identified multiple times (due to post translational modifications, different charge states and retention times), leading to repeated measurements on the peptide level.

1.2 Research questions

A handful of normalization methods have been proposed. A simple and non-exhaustive categorization could divide those methods into two groups: data-driven or model-based techniques. Quantile normalization, NOMAD [4], median polish, or CONSTANd [5] are examples of the first group. These methods scale well with the constantly growing size of proteomic datasets and do not assume any underlying statistical model. On the other hand, it is argued that one of the state-of-the-art approaches for data normalization in the context of isobaric labeling is an approach based on ANOVA models, as proposed and illustrated by Hill et al. [6] and Oberg et al. [7]. However, taking into account the fact that a dataset comprising of tens of thousands of observations is not uncommon in quantitative proteomics, fitting an ANOVA model becomes a real computational challenge. Moreover, different ways of performing data normalization may differently affect the DEA results - the last and central step of a quantitative, MS-based experiment. Therefore, the comparative analysis of the two aforementioned groups of data normalization techniques becomes an important research topic.

In particular, the goal of this thesis is to investigate the performance of two normalization methods: CONSTANd and ANOVA based approach. Since the ANOVA method relies on more advanced mathematical theory than CONSTANd does (as illustrated later in the *Methods* section), it is hypothesised that ANOVA should outperform CONSTANd. This statement will be evaluated on two real-life datasets, based on:

1. normalization capabilities, assessed with diagnostics plots (boxplot, principal component analysis (PCA) plot and heatmap),
2. differential expression analysis results, assessed by protein $\log 2$ fold change and p-value,
3. computation time of each data normalization method.

The report has been organised in the following way. *Methods* section gives a short introduction to linear mixed models framework and describes its application in the quantitative proteomics context. Moreover, CONSTANd normalization is briefly discussed and the section ends with a guide to the analysis workflow and information on the datasets analyzed in this research. *Results* section contains numerical results as well as informative data visualizations obtained from two real life datasets. In *Discussion* section, project findings are summarized and ideas for future research are presented.

2 Methods

2.1 State-of-the-art ANOVA approach

Linear mixed models introduction

It is easy to see that an ANOVA model can be considered as a special case of linear regression (a linear model including only categorical explanatory variables), and the linear modeling can be perceived as a special case of the linear mixed modeling (LMM) framework. Also, the LMM framework might be especially useful in light of the described multilevel data hierarchy above. As such, a short introduction of linear mixed models will follow.

The common definition of linear mixed model is:

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \\ \\ \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{i}), \\ \\ \mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \text{ independent} \end{cases} \tag{1}$$

where index i corresponds to the i-th subject; $\mathbf{Y}_i$ is the response variable vector; $\boldsymbol{\beta}$ and $\mathbf{b}_i$ are the fixed and random effects respectively; $\boldsymbol{\varepsilon}_i$ is the residual error. Matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}_\mathbf{i}$ describe the variance components related to the random effects and residuals errors, respectively.

An interpretation of such a model is twofold. The marginal interpretation yields that:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_\mathbf{i}), \tag{2}$$

while the hierarchical interpretation implies that the distribution of the response vector is:

$$\mathbf{Y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_\mathbf{i}) \tag{3}$$

A technical comment: within the frequentist paradigm, only the marginal model can be fitted to the data. However, the empirical Bayes technique allows estimation of the random effects $\mathbf{b}_i$, so the subject-specific profiles/trajectories can be obtained. The estimation of the fixed effects and variance components in equation 1 is done either via Maximum Likelihood (ML) or REstricted Maximum Likelihood (REML) methods. Providing exact formulas for the estimation process (which, nota bene, vary depending on the software and package used to fit mixed models) is outside of the scope of this report. Note that the fixed effects are estimated conditionally on the estimated variance components, therefore the inclusion of random effects not only affects inference for the fixed effects, but also affects their point estimates.

It can immediately be noticed that the LMM framework allows for deviations from the two assumptions of a regular linear model, namely: independence of observations and homogeneity of the residual variance. Such deviations frequently occur while analysing longitudinal, clustered or spatial data. Moreover, LMM can handle complex hierarchical structures. If there is no reason to assume any actual hierarchy in your data, a flexible

variance structure (for the random effects and the residuals) depending on categorical predictors can be specified. That enables partitioning the overall variance into the variance components corresponding to the specified grouping factors, which may also be of interest for the researcher. Again, if the hierarchical interpretation is assumed, the subject-specific profiles/trajectories/predictions can be obtained via empirical Bayes method. Finally, in the presence of variance heterogeneity and/or correlation among observational units, LMM enables for a valid statistical inference.

ANOVA based normalization and DEA

Several approaches for data normalization can be employed in quantitative MS-based proteomics, but not all of them are flexible enough to be successfully applied in an experiment with a complex design. Additionally, data-driven methods usually do not allow the explicit specification of the known sources of confounding in the normalization procedure.
These two aspects, flexibility and explicit specification of the source of confounding, may be considered the most important reasons to use an ANOVA based approach.

Similarly to the microarray setting, Hill et al., [6] propose to use the ANOVA framework for normalization and differential expression analysis of data resulting from iTRAQ experiments. In their work, the authors explain biological and experimental factors that may contribute to the overall variability in the observed reporter ions peaks. Those factors are translated and incorporated into an ANOVA model. Using their notation, the model is as follows:

$$\log_2 y_{i,j(i),c,q,l,s} = \text{intercept} + p_i + r_{i,c} + r_c + f_{j(i)} + g_{j(i),c} + v_{q,l} + b_q + \varepsilon_{i,j(i),c,q,l,s} \quad \text{(Model 1)}$$

where:

- $y_{i,j(i),c,q,l,s}$ - the observed reporter ion intensity value,
- p_i - the overall i-th protein contribution,
- $r_{i,c}$ - the i-th protein contribution in condition c,
- r_c - the overall effect of condition c,
- $f_{j(i)}$ - the overall contribution of the j-th peptide nested in the i-th protein,
- $g_{j(i),c}$ - the contribution of the j-th peptide nested in the i-th protein in condition c,
- $v_{q,l}$ - the effect of l-th quantification channel nested in q-th MS^2 run,
- b_q - the effect of the q-th MS^2 run,
- $\varepsilon_{i,j(i),c,q,l,s}$ - the model's residual capturing biological and measurement errors, on spectrum level (index s).

Note that a combination of the q and l indexes unambiguously determines the study sample.

The response variable is log transformed (the logarithm base can be specified arbitrarily) because researches usually assume the multiplicative effect of predictors on the ion intensities.

For a more detailed interpretation (also from a wet-lab perspective) of the terms included in the model, see the source article. The normalization and differential analysis steps are combined in one model, potentially leading to more efficient inference.

The authors evaluated Model 1 performance on a small dataset prepared in-house which contained eight proteins mixed with known amounts and analyzed with a single MS^2 run. Very often biological samples contain hundreds of proteins and studies are designed in such a way that multiple MS^2 runs are required.
Consequently, Oberg et al., 2008 [7] responded to the need of showcasing capabilities of the ANOVA method on a more complex dataset. In that study, 992 unique proteins and 2637 unique peptides were identified in samples from three histological subtypes of acute cardiomyopathy (the study design included six MS^2 runs and four quantification channels). At that time, Model 1 could not be fitted using available software and computers, therefore the authors proposed to split the procedure into two parts: 1) data normalization and 2) differential expression analysis. Step 1) involves a model with all terms as in Model 1 definition except the variables indexed with "c" i.e. the biological condition:

$$\log_2 y_{i,j(i),q,l,s} = \text{intercept} + p_i + f_{j(i)} + v_{q,l} + b_q + \varepsilon_{i,j(i),q,l,s}. \qquad \text{(Model 2)}$$

Then, the Model 2 residuals given by:

$$w_{i,j(i),q,l,s} = \log_2 y_{i,j(i),q,l,s} - (\widehat{\text{intercept}} + \hat{p}_i + \hat{f}_{j(i)} + \hat{v}_{q,l} + \hat{b}_q) \qquad (4)$$

are the normalized reporter ions intensities and can be treated as the response variable for step 2).

Turning now to step 2), one DEA model for all proteins simultaneously could be fitted to the normalized data:

$$w_{i,j(i),c,q,l,s} = \text{intercept} + r_c + r_{i,c} + g_{j(i),c} + \eta_{i,j(i),c,q,l,s}, \qquad \text{(Model 3)}$$

where $\eta_{i,j(i),c,q,l,s}$ represent the residual error.

However, even such a partial model may be computationally challenging (still thousands of coefficients to estimate) and not be feasible for large studies investigating complex organisms.

Another (computationally less challenging) example of step 2) could be an ANOVA model fitted to each protein separately (a protein-by-protein model):

$$w_{j,c,q,l,s} = \text{intercept} + r_c + \eta_{j,c,q,l,s} \qquad \text{(Model 4)}$$

where index i corresponding to proteins is deliberately omitted.

Two comments on the two step approach were given by Oberg et al.. First, the protein and peptide effects in step 1) (p_i and $f_{j(i)}$ terms) can be specified as random effects. As it will be shown in the following sections of this report, that remark essentially enabled the model based normalization of large proteomic datasets analyzed in this research in a reasonable amount of time. Otherwise, the fixed effects specification of the protein and peptide contributions would lead to a very time consuming model fitting process.
Second, they recommend to perform step 2) in a protein-by-protein fashion, not only because it is a much faster way to perform DEA, but also because it implies that each protein has a different variance parameter. This assumption seems to be more plausible than assuming constant variance across all proteins implied by Model 3. Therefore the protein-by-protein way of differential testing will be adopted in the analysis workflow presented in this thesis.

2.2 CONSTANd

The CONSTANd method has been developed by Maes et al., 2016 [5], and is a fast and data-driven normalization algorithm applicable to, among others, TMT/iTRAQ studies. The input for this method is a matrix **A** resulting from a single MS^2 run, where the rows correspond to peptides and the columns are representing quantification channels. Note that the method assumes only one quantification value per peptide in the sample, therefore the peptide repeated measurements (due to PTM, charge state and RT) must be aggregated beforehand.

The procedure is built upon two constraints. First, the measured reporter ion intensities are of relative nature, thus they should be interpreted as a percentage (summing to one). Second, usually a significant effort is put in by a wet-lab analyst in ensuring that the protein concentration is the same across samples within a single MS^2 run. Therefore, the problem at hand is to find a new matrix **K** that deviates the least from the original quantification matrix **A**, subjected to the aforementioned constraints. This problem was successfully tackled with a simple and elegant procedure originating from the field of econometrics, called RAS, which falls within a more general framework of Iterative Proportional Fitting procedure (IPFP).
It turns out that the solution is given by:

$$\mathbf{K} = \mathbf{R} \times \mathbf{A} \times \mathbf{S},$$

where **R** and **S** are diagonal matrices containing the inverse of row and column sums, respectively.

Advantages of CONSTANd are manifold. First, the normalization is performed for each MS^2 run independently, so there is no need to wait until all samples are analyzed (on-line mode). Second, CONSTANd does not require any reference samples or spiked-in proteins to obtain correct global normalization (i.e. across all MS^2 runs in the experiment). Moreover, the column constraint specified in CONSTANd allows for elimination of the bias across quantification channels, making the quantification values comparable within a MS^2 run (intra-run comparability). When the row constraint is added on top of that, the peptide quantification values become comparable also between MS^2 runs (inter-run comparability).

2.3 Analysis workflow

Figure 6 shows the flow of the planned analysis steps, steps that are applicable to any proteomic dataset treated with isobaric labeling. Importantly, in the beginning of the analysis, the quantification data resulting from multiple MS^2 runs are stored in separate data frames.

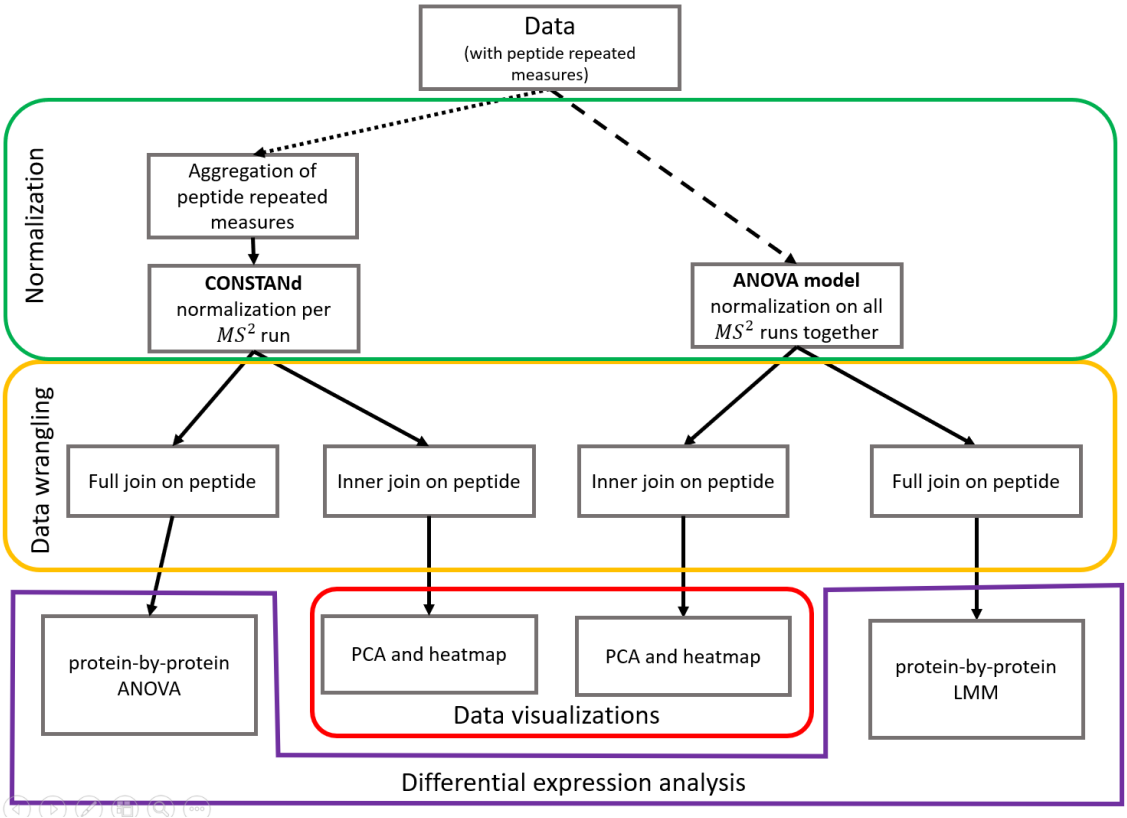

Figure 6: An overview of the planned analysis. On the left hand side, CONSTANd based analysis is shown, and it involves CONSTANd normalization followed by protein-by-protein ANOVA model fitting. On the right side, ANOVA based analysis is depicted, and it consists of data normalization with ANOVA model fitted to unaeggregated data, followed by a protein-by-protein liner mixed model fitting. In both analyses, data normalization is visually assessed based on PCA plot and heatmap.

CONSTANd based analysis

Starting from the left branch of the diagram (the dotted arrow on the left), repeated measurements on the peptide level must be aggregated prior to applying CONSTANd. The aggregation involves selection of the best representative measurement per peptide. "The best" means having the highest score given by the scoring algorithm used in the database search (for instance Mascot), and if there is a tie, having the highest sum of intensities across all quantification channels. In the next step CONSTANd is used to normalize each MS^2 run data frame.
The quality of the performed normalization is assessed with the following visualizations: boxplot, PCA plot and heatmap. The input data for PCA plot and heatmap is a subset of peptides detected in each MS^2 run (denoted as "Inner join on peptide" in Figure 6). This is because PCA and heatmap techniques cannot handle missing values. When it comes to DEA, that step is performed on all peptides across all MS^2 runs (denoted as "Full join on peptide" in Figure 6).

DEA includes the estimation of protein $\log_2$ fold change (FC) :

$$\log_2 \mathrm{FC}_i = \log_2 \frac{\bar{y}_{i,c}}{\bar{y}_{i,ref}} \tag{5}$$

where i indicates the protein, and $\bar{y}_{i,c}$, $\bar{y}_{i,ref}$ correspond to the mean of the all normal-

ized peptide quantification values (from all samples and peptides) belonging to the group denoted by the index "c" and the reference group, respectively.

P-values, indicating which proteins demonstrate significantly different expression patterns between certain groups of samples, are obtained by fitting an ANOVA model. However, one additional step is required in order not to violate one of the linear model assumptions - independence of observations. For a given protein, the quantifications values corresponding to multiple peptides of that protein are averaged within each sample. This step has been introduced to mimic the QCQuan analytical workflow for iTRAQ/TMT studies described in Van Houtven et al., 2019 [8]. Thus, the following protein-by-protein ANOVA model is fitted:

$$z_{q,l} = \text{intercept} + r_c + \eta_{q,l} \qquad \text{(Model 5)}$$

where q and l correspond to the experimental run and quantification channel, $z_{q,l}$ is the average of the normalized peptide quantification values from the sample analyzed in the q-th run and l-th channel. Moreover, r_c is the group indicator of the experimental samples and $\eta_{q,l}$ is the residual error. Note that the $z_{q,l}$ values are not log transformed prior to the model fitting. Only p-value yielded by Model 5 will be investigated, while $\log_2$ FC is computed according to equation 5.

ANOVA based analysis

The right side of the diagram displayed (the dashed arrow on the right) in Figure 6 pertains to the ANOVA based analysis. The first step is to combine data frames corresponding to each MS^2 run into one data frame. In contrast to CONSTANd, the repeated measurements on the peptide level (due to PTMs, charge state and RT) are not aggregated and kept in the dataset. Then, the two step procedure (described by Oberg) is performed. A normalization model similar to Model 2 is fitted, and its residuals are considered as the normalized quantification values (see equation 4). An exact specification of the normalization model depends on the analysis design, therefore a more detailed description of the model terms is included in the *Results* section. Afterwards, PCA plot and heatmap are created based on the subset of peptides detected in each MS^2 run and the corresponding normalized quantification values.

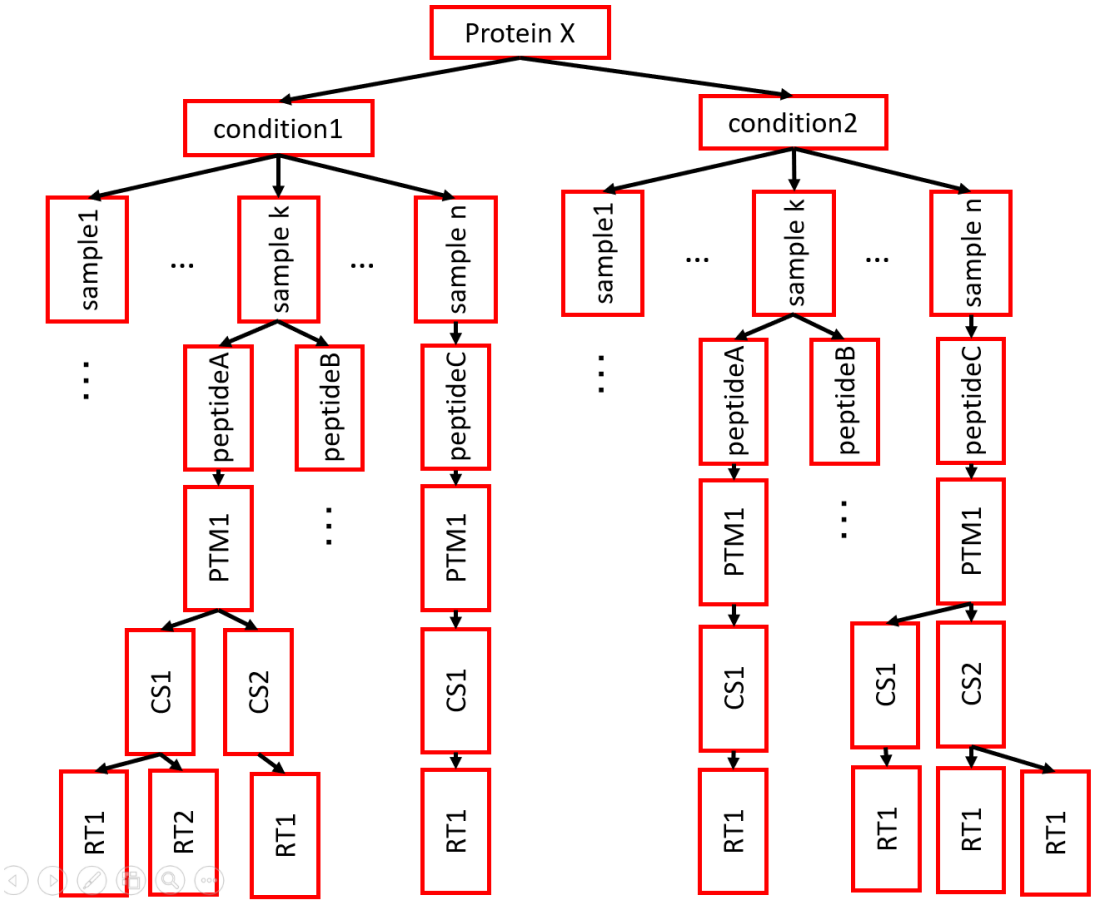

Figure 7: A potential data hierarchy of a certain protein viewed from differential expression analysis perspective, in which significance testing is performed on a protein-by-protein basis. One protein can be present in multiple biological samples taken from several treatment arms or other biological groups of interest (here condition 1 and condition 2). Peptide quantification values within a sample may be considered as the repeated measurements of the protein and that relationship should be accounted for in the statistical analysis, for example, by linear mixed models. However, some proteins detected in some samples may be represented by one peptide quantification value only, posing a problem with fitting a mixed model.

As presented in Figure 7, a multilevel hierarchy may exist in the data for a given protein (here "Protein X"), indicating that the linear mixed models framework might be useful in such a setting. Therefore, the DEA consists of a protein-by-protein linear mixed model fitting to the residuals (the normalized values) yielding model-based $\log_2$FC estimates (the model's coefficient) and p-values. A protein specific model, similar to Model 4, is given by:

$$w_{j,c,q,l,s} = \text{intercept} + r_c + d_{q,l} + \varepsilon_{j,c,q,l,s} \qquad \text{(Model 6)}$$

Note that the i index corresponding to the proteins is discarded as in this procedure a separate model is computed for each protein. The $d_{q,l}$ term is the sample random effect accounting for the fact that the peptide quantification values within the sample may be correlated. Note, however, that some proteins may be represented by one peptide only and that peptide may have no repeated measurements (due to PTMs, charge state or RT). In such a scenario, variance of $d_{q,l}$ may be estimated as zero and that would lead to a

linear model with fixed effects only.

2.4 Datasets

The first dataset (called `MICE data` throughout this text), introduced in Maes et al. [5] consists of protein expression data from 6 mice with orthotopic non-small cell lung adenocarcinomas. From each mouse 3 samples were isolated, one with tumor-associated macrophages (TAM), one with peritoneal macrophages (in-vivo control, PM) and one with bone marrow to produce macrophages (in-vitro control, BM), leading to 18 biological samples in total. The samples then were treated with TMT 6-plex isobaric labels and analyzed with three MS^2 runs. The allocation of samples into MS^2 runs and TMT labels was done in a block-randomized fashion (blocks are TMT labels, allocated "treatment" is TAM, PM and BM), in such a way that only two mice are present in each MS run. Figure 8 displays the design. The goal of this study was to compare the three biological conditions.

ID/ label	TMT⁶-126	TMT⁶-127	TMT⁶-128	TMT⁶-129	TMT⁶-130	TMT⁶-131
tmt1	BM3	PM3	TAM4	BM4	TAM3	PM4
tmt2	PM5	TAM6	BM5	TAM5	PM6	BM6
tmt3	TAM1	BM1	PM1	PM2	BM2	TAM2

Figure 8: The design of MICE dataset. Graph based on Table 1 in Maes et al., 2016 [5].

Next, raw spectra were processed with Proteome Discoverer version 2.0 using MASCOT/Sequest HT (MASCOT as a master database search engine and Sequest as a secondary search engine) with 5% false discovery rate (FDR) control based on the reverse target-decoy database approach. The data was also filtered as described in Van Houtven et al., 2019 [8] (the "Processing step"). The filtering included:

1. Removing PSMs with "Confidence level" (if available) worse than "Medium" or with "Isolation Interference" [%] level (if available) higher than 30.
2. For each MS^2 run, only peptides present (i.e. having non-missing intensity value) at least once in each biological condition were kept. For instance, peptide X from the first TMT run with missing data for channels 126 and 129 would be omitted from further analysis.

The filtering steps above resulted in a dataset with 1257 unique proteins, 3666 unique peptides, and a 3666×18 quantification matrix. In Figure 9, it can be seen that approximately only one third of proteins has been identified in all three experiments, whereas this percentage is even lower for peptides (only about 24%). This shows that MS based studies may suffer from low reproducibility of findings.

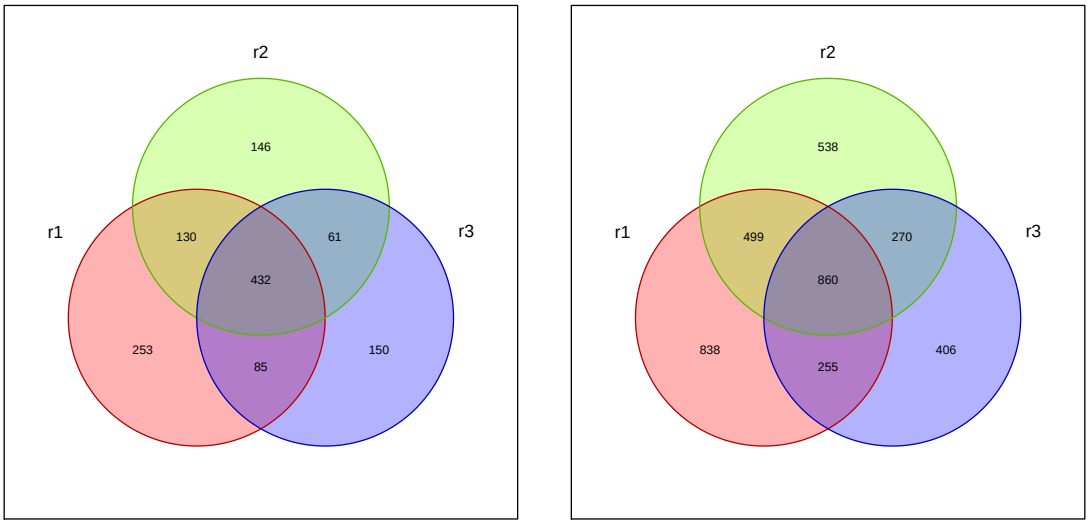

Figure 9: Venn diagram of the MICE data. Left panel shows unique proteins, right panel presents unique peptides identified across three MS^2 runs (r1, r2, r3).

The second dataset (called `Erwinia data` throughout this text) was introduced in Gatto et al. [9] and, as the author says, "in this TMT 6-plex [23] experiment, four exogenous proteins were spiked into an equimolar Erwinia carotovora lysate with varying proportions in each channel of quantitation; yeast enolase (ENO) at 10:5:2.5:1:2.5:10, bovine serum albumin (BSA) at 1:2.5:5:10:5:1, rabbit glycogen phosphorylase (PHO) at 2:2:2:2:1:1 and bovine cytochrome C (CYT) at 1:1:1:1:1:2". As such, this spiked-in dataset can be used for benchmarking the ANOVA method approaches by performing differential expression analysis. The data is publicly available at `http://www.proteomexchange.org/` under `PXD000001` accession number.

1305 unique proteins and 3117 unique peptides were identified in the dataset. As this is a single TMT experiment, without any explicitly defined factors that could be considered as a "treatment arm" or "biological condition", the TMT-129 channel will be treated as a reference sample.

The two following comments regard to both datasets. The raw data include the peptide repeated measurements and this may allow checking whether those additional observations bring useful information or rather add more noise to downstream analysis. Also, proteins with only one peptide identification, sometimes called "one-hit wonders", were not disregarded from any of the statistical analyses presented in this report.

3 Results

All results were obtained using `R` software, version 3.6.1. The following functions and R packages were utilized: `lmer` implementation of linear mixed models from package `lme4`, `CONSTANd` python script run in R with `reticulate` package.

3.1 Mice data

Prior to applying the two step procedure described by Oberg, two attempts were made to fit Model 1 to the MICE data. A reason to do that is the fact that fitting one large model instead of the two step approach could lead to more efficient statistical inference. However, approximately 35 hours after starting the computations, each time an error about insufficient RAM resources was encountered. Not surprisingly because on the MICE data, Model 1 implies the estimation of about 19 000 coefficients. Even if DEA is not performed on the peptide level, meaning that $g_{j(i),c}$ term is omitted, such a model is still computationally too complex to fit on a personal computer. Thus, the only option is to resort to the aforementioned two step procedure.

Furthermore, it was found that treating protein and peptide effects $p_i, f_{j(i)}$ in Model 2 (the normalization model) as random effects speeds up the fitting process significantly, decreasing the running time from hours to about one minute. This is due to estimating only covariance parameters of the random effect distributions instead of estimating thousands of protein and peptide fixed effects. Therefore, from this point the expression "ANOVA method" for the state-of-the-art approach introduced in Oberg will be replaced with "LMM".

Another important finding observed on the MICE data is that it should be allowed to vary the peptide random effect $f_{j(i)}$ together with MS^2 run. In general, MS^2 is such a complex physico-chemical process that differences in intensities across MS^2 experiments for the same peptide can be anticipated. One of the possible explanations of those differences might be different ionization efficiencies across multiple runs of a mass spectrometer. As such, a good normalization model should account for that phenomenon, for example, by including the $f_{j(i)}$ term specified as in Model 7.

Hence, the following normalization model was fitted on the MICE data (data that still includes the peptide repeated measurements):

$$\begin{aligned} \log_2 y_{i,j(i),q,l,s} = \text{intercept} + p_i + f_{j(i)} + v_{q,l} + b_q + \varepsilon_{i,j(i),q,l,s}, \\ p_i \sim N(0, \sigma_p^2), \\ f_{j(i)} \sim N(\mathbf{0}, \boldsymbol{C}), \\ \boldsymbol{C} = \begin{pmatrix} c_{11} & & \\ c_{21} & c_{22} & \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \end{aligned} \qquad \text{(Model 7)}$$

where $\boldsymbol{C}$ is a covariance matrix and its elements c_{ij} specify the variance-covariance relationship of the peptide random effects across MS^2 runs.

Model 7 will be denoted as "LMM" in the following text, figures and tables.

A similar, but more simple model:

$$\begin{aligned} \log_2 y_{i,j(i),q,l,s} = \text{intercept} + p_i + f_{j(i)} + v_{q,l} + b_q + \varepsilon_{i,j(i),q,l,s} \\ p_i \sim N(0, \sigma_p^2), \\ f_{j(i)} \sim N(0, \sigma_f^2), \end{aligned} \qquad \text{(Model 8)}$$

will be denoted as "LMM no interaction" (no interaction refers to the peptide random effect being constant across runs for the same peptide) in the following text, figures and tables.

A few other normalization model specifications were considered in addition to Models Model 7 and Model 8, for instance, a model treating peptides as random effects nested within runs or a model having both: protein-by-run and peptide-by-run random effects at the same time. Although resulting data for those models are not reported in this thesis report, Model 7 was informally selected based on AIC and BIC criteria and its clear interpretation. Also note that no formal model diagnostics were performed on Model 7, because there is no interest in conducting formal statistical inference on the technical effects $p_i, f_{j(i)}, v_{q,l}, b_q$. The only objective is to normalize the data correctly, and that can be assessed with the normalization data visualizations mentioned above.

3.1.1 Data normalization and computation times

As mentioned previously, normalization quality of CONSTANd and the LMM method will be assessed via boxplot, PCA plot and heatmap. In the following plots, the input data for the "Before normalization" panel included no repeated measurements on the peptide level as they were aggregated using "the best representative" approach. In the "CONSTANd" panel, the normalized quantifications values were used as is. Inputs for the "LMM no interaction" and "LMM" panels are the averages (over PTMs, charge state and RT) of the model-based normalized quantification values. This was done in order to aggregate the peptide repeated measurements and obtain one quantification value per peptide within a sample (required by PCA plot and heatmap).

Figure 10 depicts the distribution of the reporter ion intensities, before and after application of the normalization procedures. It can be noticed that peptide distributions are well aligned around the overall mean intensities in the entire study (the red line) after applying any normalization approach.

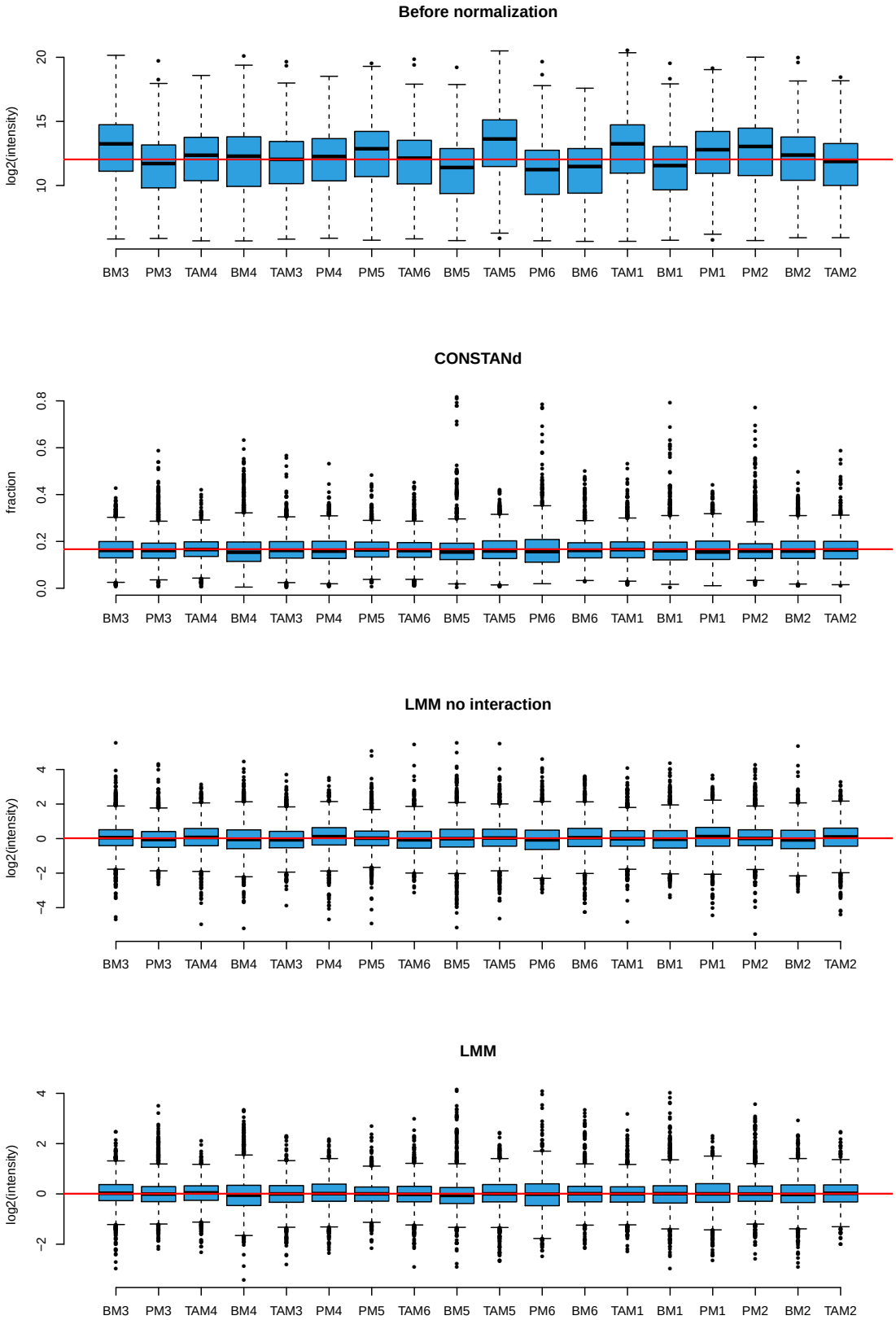

Figure 10: Boxplots of peptide reporter ion abundance. Samples are presented in line with the MS^2 run (three series of six quantification channels). The red line represents the mean value of peptide intensities across all MS^2 runs. The y-axis label on the CONSTANd plot refers to fraction of the total peptide's intensity within a MS^2 run. All normalization approaches result in similar boxplot profiles.

Next, PCA is one of several dimension reduction techniques, which sequentially searches for a new direction in the data (i.e. a linear combination of available variables, here peptides), such that the projection onto the new coordinate has the largest variance (i.e. explain the most of observed variance). Then, the second direction with the largest variance and orthogonal to the first coordinate found in the previous step, is calculated, and so on. These new coordinates are called principal components (PC). In practice, the two first principal components are often used to create a scatter plot and to visually inspect the data at hand.

Heatmap, or hierarchical clustering (HC) plot is another technique useful in a quantitative proteomics context. It is based on similarities (distance) between observations and it starts with treating each observation as a single cluster. In the next steps, similar clusters are merged together, until all observations end up in a one big set. It is the end user who decides when to stop the merging based on (dis)similarity of clusters. Moreover, a few ways of how to assess (dis)similarity between any two clusters exist, but in this application, average distance (i.e. the average of distances between every two data points from both clusters) is used.

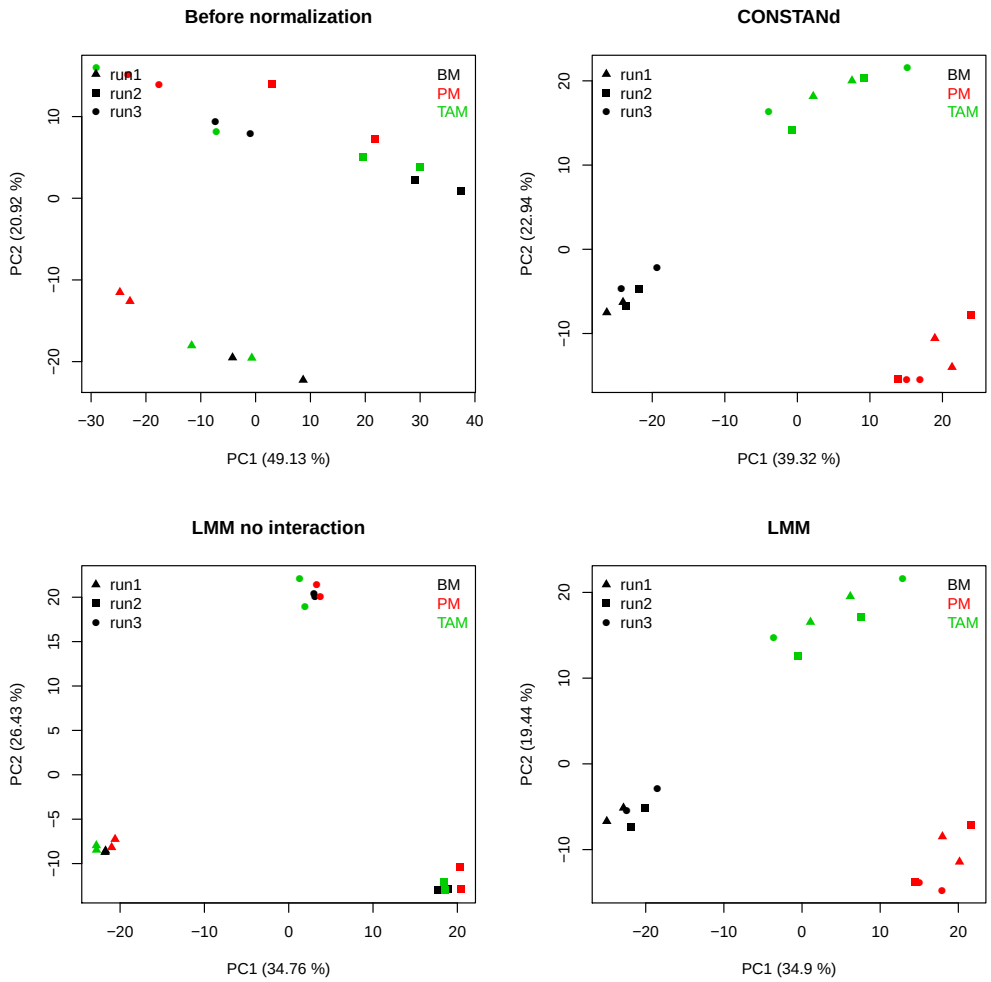

Figure 11: PCA plot based on the first two principal components (PC1 and PC2). Before normalization, biological samples are rather clustered according to MS^2 run, while the CONSTANd and LMM normalization methods yielded clusters based on biological conditions. The LMM normalization (Model 8) without the peptide-by-run interaction grouped the samples in line with MS^2 run, emphasising importance of such an interaction term in the case of multiple MS^2 run experiment.

Note that input data for PCA plots and heatmaps take form of a quantification matrix, where rows correspond to samples (observations) and columns represent peptides (features). Moreover, the techniques cannot handle missing data in any way, so first, peptides not identified in all three experiments are removed, and second, even if a peptide is detected in all three runs, still for some quantification channels missing values may occur. These missing values are replaced with zeros. This is a quite reasonable (although not the optimal) approach, because most likely those missing ion intensity values were below the level of detection of a mass spectrometer.

The results from the PCA analysis on the unnormalized dataset, shown in Figure 11, confirm that observations are grouped based on a spectrometer run (points with the same shape lie closely to each other) rather than on biology (clusters of points with mixed colours). Also, there is quite a lot of variability within the experimental runs, suggesting there are other factors masking the biological effect. On the other hand, CONSTANd and LMM managed to remove technical effects and group the samples according to the

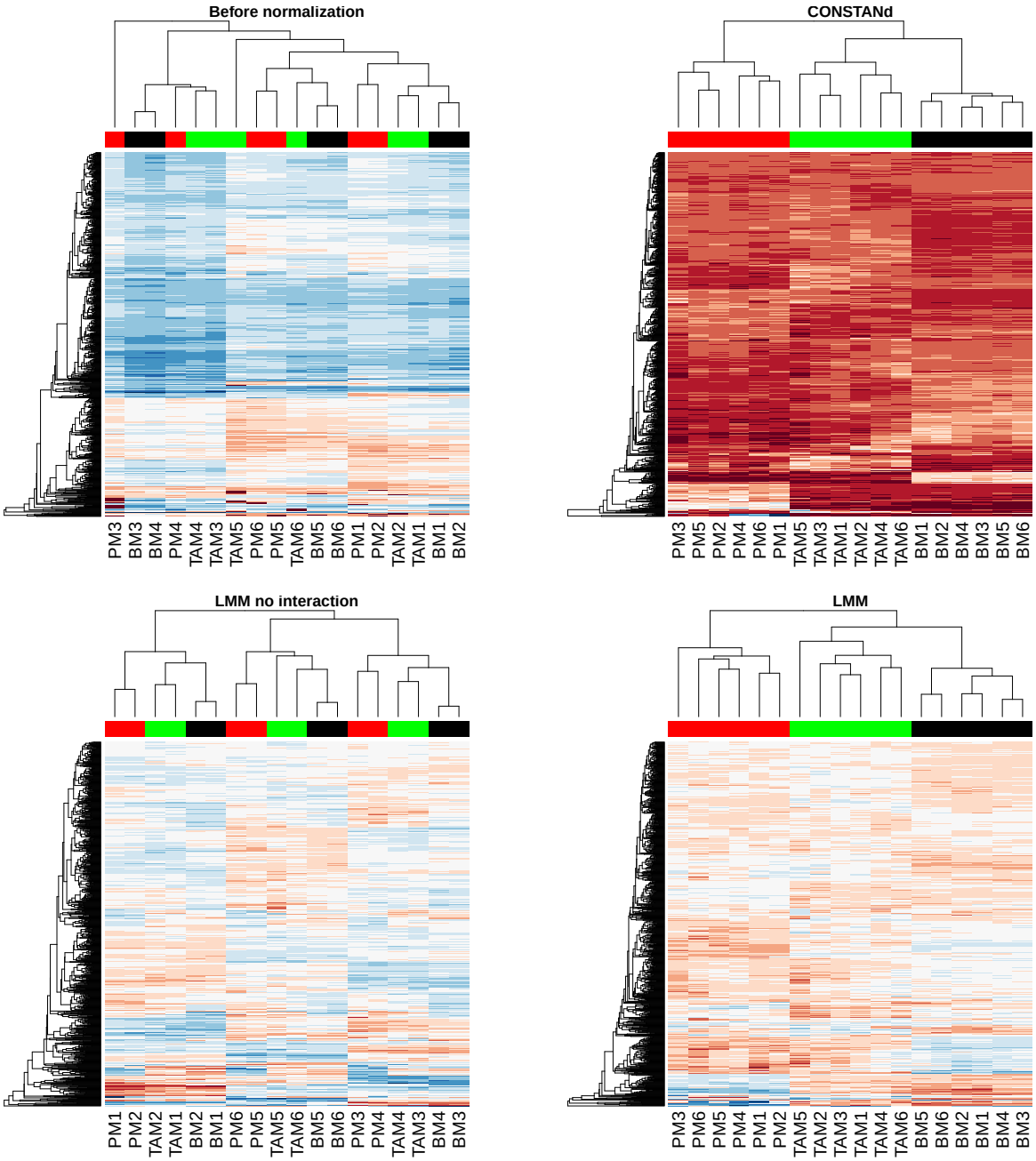

Figure 12: Heatmaps based on hierarchical clustering using the average linkage method. Column clusters correspond to the study samples. Six consecutive samples of the same condition, also indicated with black, red and green horizontal side bars, demonstrate clustering driven by the biological effect. Similarly to PCA plots, only the CONSTANd and LMM methods yielded clusters according to biology.

analyzed biological conditions (the same colours within the clusters). As communicated previously, the "LMM no interaction" based normalization organized samples according to the mass spectrometer run.

Heatmaps shown in Figure 12 yield similar conclusions as the PCA plots. The most important piece of information in these graphs are the labels on the x-axis. If normalization is successful, this will be reflected in subsequent groups (of size equal to six) of samples isolated from the same biological condition (e.g. TAM5, TAM2, TAM3, TAM1, TAM4, TAM6). In the "Before normalization" panel samples are grouped in line with the mass spectrometer run. It seems that "LMM no interaction" allocated the samples according to the biology within the MS^2 run. In contrast, CONSTANd and the LMM approach correctly put together samples according to their cell type.

In summary, these results show that both methods: CONSTANd and the LMM, are able to remove technical variability inherent to the experimental run and quantification channels. Additionally, at least in the case of MICE data, generated from a multiple MS^2 runs study, some kind of interaction of peptide and run effects was required to achieve the correct normalization with the linear mixed model. That resembles the CONSTANd's row constraint, which brings the peptide quantification values from multiple experiments into the same scale and makes them comparable between different runs of the mass spectrometer.

To better understand that effect, take a close look at one of the peptides identified in the all three experiments. Consider a peptide given by LLQDSVDFSLADAINTEFK sequence, a building block of a protein described with P20152 master accession number. In total, 476 quantification values were recorded for that peptide and are presented in Figure 13. Boxplots indicate that indeed, only when the peptide-by-run interaction is specified in the model, the peptide distributions across the experimental runs are better aligned. The same can be noticed in the bottom panels, where a more granular view on the peptide quantification values (sorted by MS^2 run) is provided. The blue lines, obtained with the LOESS smoother, become more horizontal after adding the interaction term.

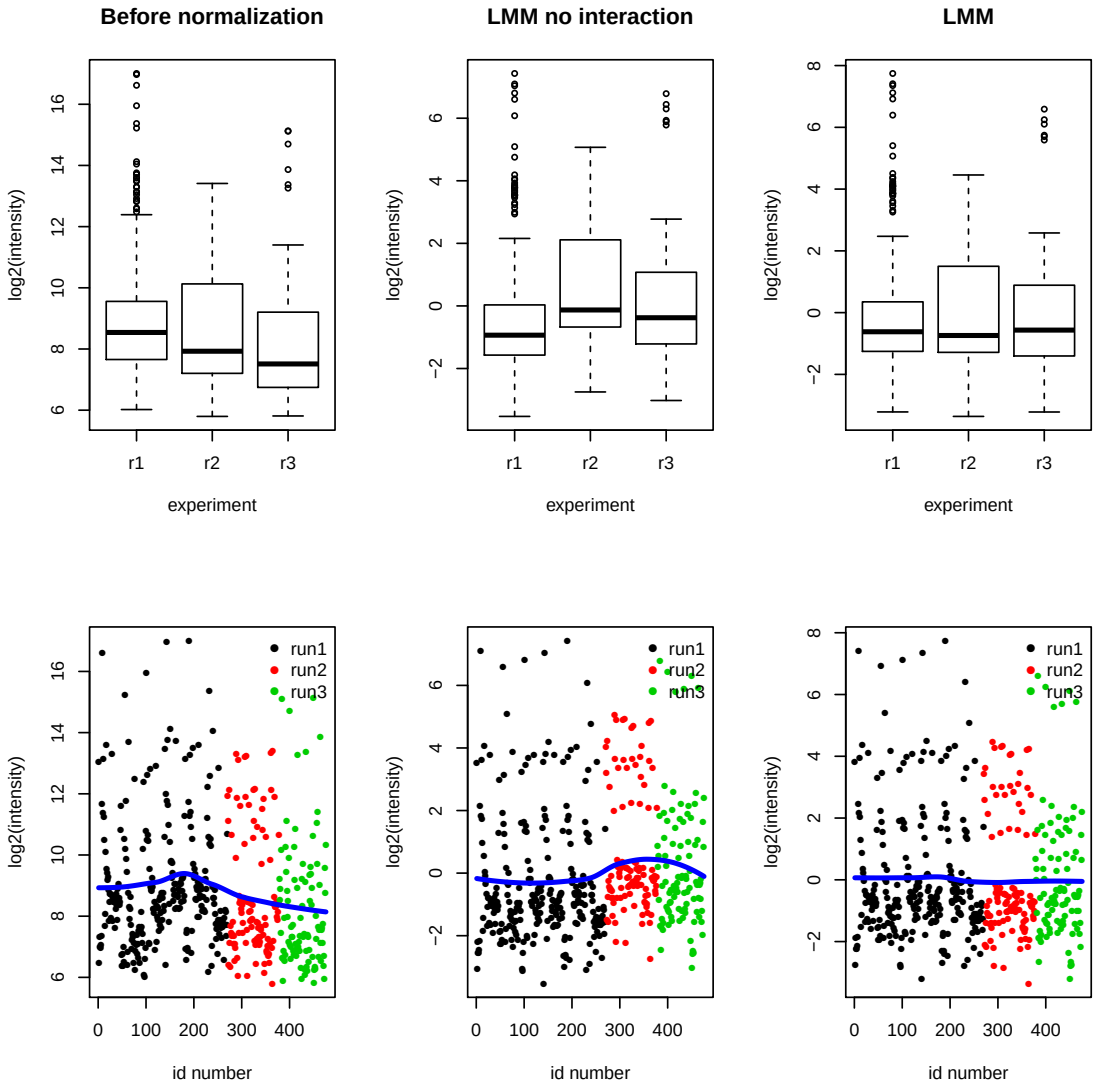

Figure 13: An investigation of the run by peptide effect based on an example sequence (LLQDSVDFSLADAINTEFK). Top panels show the peptide distributions across multiple runs. Bottom panels display the peptide quantification values sorted by MS^2 run. The blue lines were obtained using LOESS smoother. Both visualizations, boxplots and scatter plots, show that peptide-by-run effect is needed to make the normalized values comparable across multiple MS^2 runs.

Moreover, the estimated variances of the $f_{j(i)}$ random effect in Model 7 in the first, second and third run, are: 0.408, 0.355, 0.362, respectively. The correlation matrix was estimated as:

$$C = \begin{pmatrix} 1 & & \\ 0.82 & 1 & \\ 0.73 & 0.83 & 1 \end{pmatrix}$$

Hence, a strong, positive correlation can be observed. Small differences between the estimated variances and correlations might suggest trying other specifications of the covariance matrix, but as presented in PCA and HC plots, the peptide-by-run random effect plays

an important role in proper data normalization. Further investigation of the covariance structure was outside of the scope of this research.

Computation time

The computational aspects of the presented methods are enclosed in Table 1. All three normalization approaches are fast, with CONSTANd being much faster as compared to the other methods. As previously mentioned, running time of a normalization model with fixed effects only would be larger by a few orders of magnitude as compared to random effects model.

Method	CONSTANd	LMM no interaction	LMM
Median time	0,033 sec	12,82 sec	56,83 sec

Table 1: Median computation time of 10 repetitions of applying different normalization methods on MICE data on a HP laptop with Intel i5-6200U CPU@2,5 GHz 2.30Ghz and 16 GB RAM. CONSTANd is three orders of magnitude faster than LMM based normalization.

3.1.2 DEA

We start this subsection with the presentation of the numbers of DEA proteins when treating the BM samples as the reference group (Table 2).

Method	Raw data	CONSTANd	LMM no interaction	LMM
TAM vs BM	0.91	34.66	13.56	28.04
PM vs BM	7.2	41.44	21.67	35.65

Table 2: Percentage of DEA proteins after applying B-H multiple testing correction at 5% FDR level.

Not surprisingly, when the unnormalized data is analyzed as in the LMM approach (excluding the LMM based normalization), the percentage of significant changes in protein abundance is low. We can also observe lower DEA rates in the case of "LMM no interaction" analysis. This is most likely caused by inability to remove the technical effects of MS^2 runs as demonstrated in PCA plot (Figure 11) or HC plot (Figure 12). The highest DEA rates correspond to the CONSTANd analysis, which in theory at least, should be methodologically less complex and more conservative method.

In order to save space, from this moment only the TAM vs BM comparison results will be presented.

Very often in differential expression analyses based on "omics" data, the analyzed genes or proteins are ranked according to the p-value of a significance test, $\log_2$ FC or some other statistical measure/score, and those "top" molecules are graphically presented. Figure 14 shows top 20 differential proteins (as judged by Benjamini-Hochberg adjusted p-value) obtained from the CONSTANd and the "LMM" analysis. Endpoints of each line correspond to results obtained either from the CONSTANd (squares) or the LMM analysis (filled circles), while the line itself illustrates the change between the outcomes. The colours indicate whether a protein is only included in the top 20 protein list resulting from CONSTANd analysis (in blue); or only in the LMM top 20 list (in green); or in the intersection of the two lists (in orange). It can be noticed that in the case of top 20 proteins of CONSTANd

and LMM approaches, $\log_2$ FC estimates stay the same (approximately vertical lines), but the p-values resulting from the LMM analysis are smaller (larger $-\log_2$(p-value)). Moreover, the overlap of the top 20 lists is equal to 13 proteins. Additionally, the entries of the protein rankings (all proteins, not only top 20) based on the CONSTANd and LMM analyses correlates moderately well, because Spearman's and Kendall' rank correlation coefficients are equal to 0.70 and 0.87, respectively.

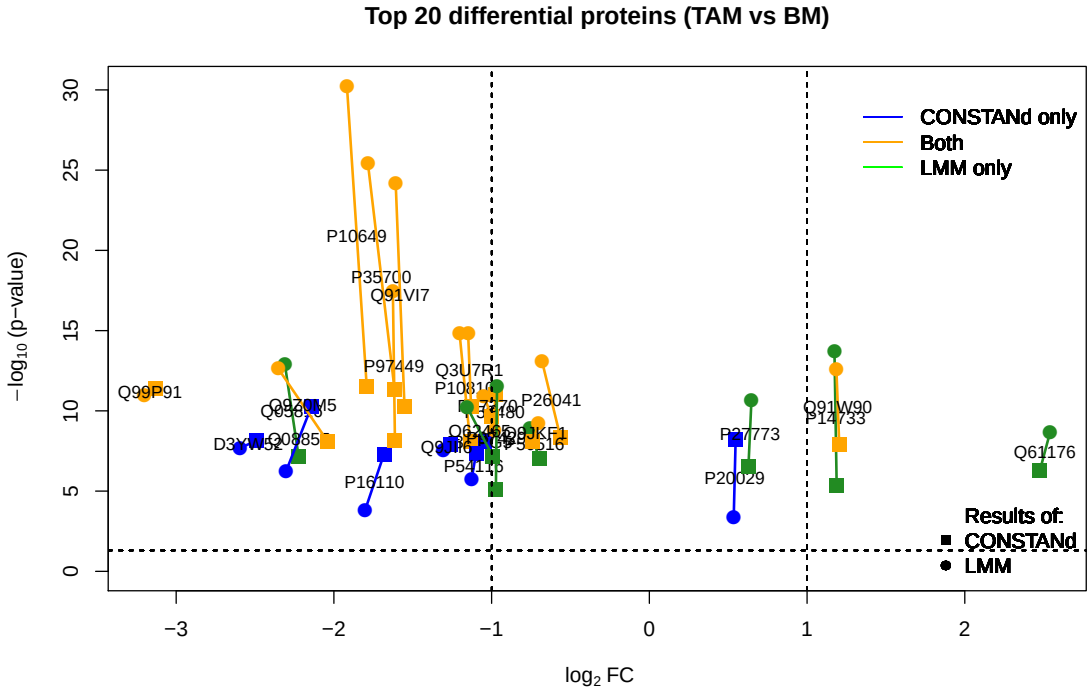

Figure 14: A volcano plot showing the top 20 differential proteins lists generated by the CONSTANd and LMM analyses. The dotted horizontal and vertical lines represent p-value equal to 0.05 and fold change of -2 and 2, respectively. Endpoints of each line correspond to the results obtained either from the CONSTANd (squares) or the LMM analysis (filled circles), while the line itself illustrates the change between the outcomes. The colours indicate whether a protein is only included in the top 20 protein list resulting from CONSTANd analysis (in blue); or only in the LMM top 20 list (in green); or in the intersection of the two lists (in orange). The overlap of top 20 differential proteins is moderate and equal to 13 proteins. The estimates of $\log_2$ FC given by the two methods are similar, while discrepancies in p-values are somewhat more pronounced.

The next graph (Figure 15) also focuses on comparing the two methods, but from a broader perspective, showing not only top 20 differential proteins, but all the analyzed proteins. Starting from the right panel, the conclusion about $\log_2$ FC based on the "top 20 plot" still holds, as the points lie closely to the identity line. Regarding p-values, the opposite behaviour can be spotted as compared to Figure 14, where $-\log_{10}$ p-value generated by the "LMM" are larger for the majority of the top proteins. Here, 830 and 379 proteins' p-values are situated above and below the identity line, respectively, meaning that "LMM" seems to be more conservative than the CONSTANd analysis. This result is in line with Table 2 and it means that either the LMM approach indeed estimates the variability present in data better due to the addition of random effects, or using full data hierarchy (i.e not aggregating the peptide repeated measurements) adds more noise instead of useful information.

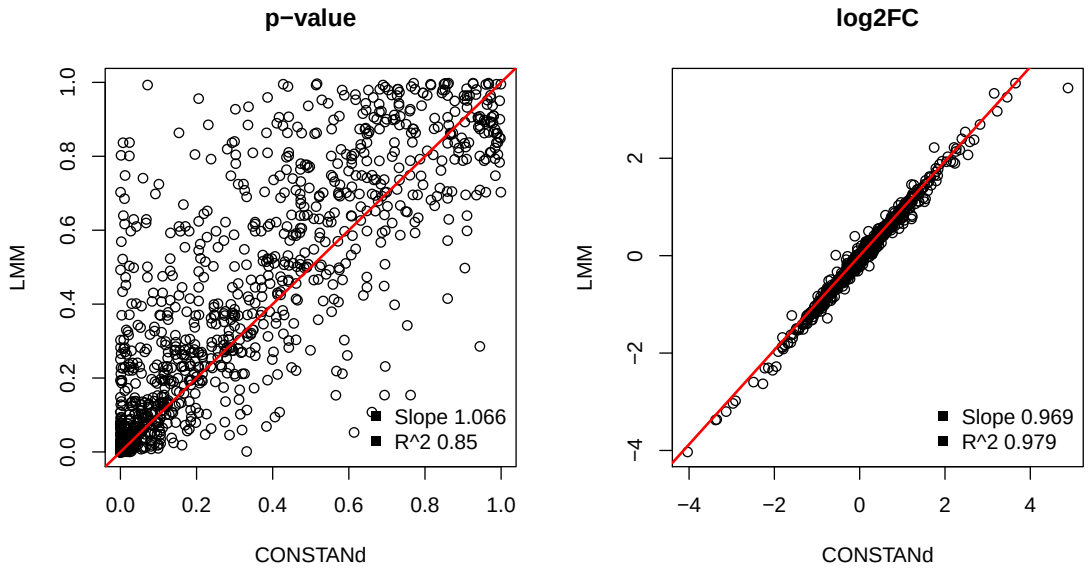

Figure 15: A comparison of DEA results yielded by the CONSTANd (x-axis) and LMM analsyses (y-axis). The identity line (y=x) is marked with red colour. For 830 proteins the LMM analysis yielded larger p-values, while the opposite is true for 379 proteins.

To check this hypothesis, the entire "LMM" analysis was repeated. This time however, the peptide repeated measurements in the raw data were aggregated with "the best representative" approach followed by data normalization with Model 7. Then the normalized quantification values were used as a response variable in the DEA analysis (Model 6). This analysis is denoted as "LMM aggregation" and results are displayed in Table 3 and in Figure 16.

	CONSTANd	LMM	LMM aggregation
TAM vs BM	34.66	28.04	35.65
PM vs BM	41.44	35.65	43.51
Spearman's corr	–	0.70	0.83
Kendall's corr	–	0.87	0.95

Table 3: Comparison of DEA results obtained from CONSTANd based analysis, LMM and LMM aggregation. Spearman's and Kendall's rank correlation were calculated by comparing p-values generated by CONSTANd based analysis and a model based approach ("LMM" or "LMM aggregation").

The "LMM aggregation" approach yielded 35.65% and 43.51% of differential proteins in TAM vs BM and PM vs BM testing, respectively, so these values are close to the CONSTANd based analysis results. Now Spearman's and Kendall' rank correlation coefficients are 0.83 and 0.95, respectively, which means better agreement between the protein rankings generated by CONSTANd and the "LMM aggregation" analyses.

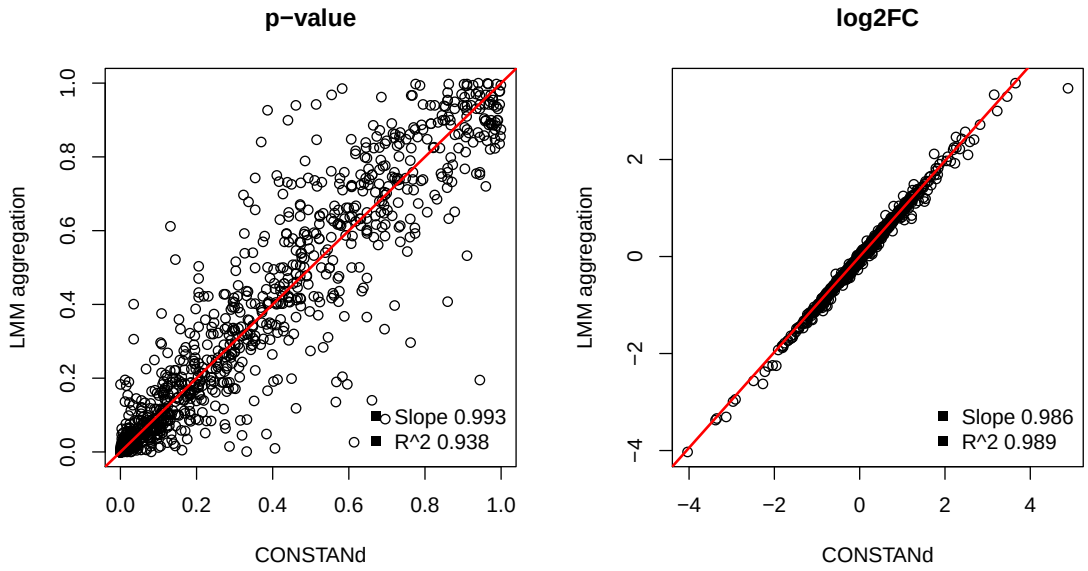

Figure 16: Results of the repeated DEA analysis based on linear mixed models fitted to data with aggregated peptide repeated measurements ("LMM aggregation"). The identity line (y=x) is marked with red colour. This time, for 623 proteins the LMM analysis yielded larger p-values, while the opposite is true for 586 proteins. Overall, the differences in p-values generated by the two methods seem to be rather stochastic in nature.

Furthermore, on the left panel of Figure 16, 623 and 586 proteins are located above and below the identity line, respectively, so the previously discussed disproportion became much smaller. Also, p-values are now more closely dispersed around the identity line as compared to the "LMM" analysis, where p-values are rather grouped in the upper-left triangle of the graph. All this taken into account suggests that it is likely not the normalization or DEA method that causes the observed differences in detecting significant proteins, but rather the type of input data (aggregated versus unaggregated).

3.2 Erwinia data

As highlighted in section 2.4, this is rather a technical dataset consisting of a single MS^2 run and six quantification channels. The channel number four was (arbitrarily) selected as the reference sample. Since there are no actual biological conditions or treatment arms in this experiment and consequently no multiple biological replicates (samples), only $\log_2$ FC can be estimated, without carrying out statistical inference (p-value).

3.2.1 Data normalization

As in the case of MICE data, CONSTANd normalization was applied to Erwinia data, as well as the following LMM normalization model:

$$\begin{aligned} \log_2 y_{i,j(i),s,l} = \text{intercept} + p_i + f_{j(i)} + v_l + \varepsilon_{i,j(i),s,l} \\ p_i \sim N(0, \sigma_p^2), \\ f_{j(i)} \sim N(0, \sigma_f^2), \end{aligned} \qquad \text{(Model 9)}$$

where now v_l corresponds to the l-th quantification channel effect (considered as a "treatment arm" in DEA later on). Note that Model 9 does not include a peptide-by-run random

effect since only one MS^2 run was performed in this study. The Model 9 residuals:

$$w_{i,j(i),s,l} = \log_2 y_{i,j(i),s,l} - (\widehat{\text{intercept}} + \hat{p}_i + \hat{f}_{j(i)} + \hat{v}_l) \tag{6}$$

will serve as an input for DEA.

Keep in mind that prior to applying CONSTANd normalization, aggregation of the peptide repeated measurements (see Figure 6) is obligatory, hence "the best representative" aggregation was also applied on Erwinia data. The aggregation of the peptide repeated measurements is not needed prior to fitting Model 9, therefore the model's residuals still reflect the peptide repeated measurements.

The normalization results can be inspected in Figure 17. After applying both normalization methods peptide distributions are better aligned around the overall mean reporter ion intensity. Other normalization visualizations like PCA plot or heatmap are not very meaningful in the case of MICE data because only one MS^2 run was performed and the study samples do not correspond to any biological condition.

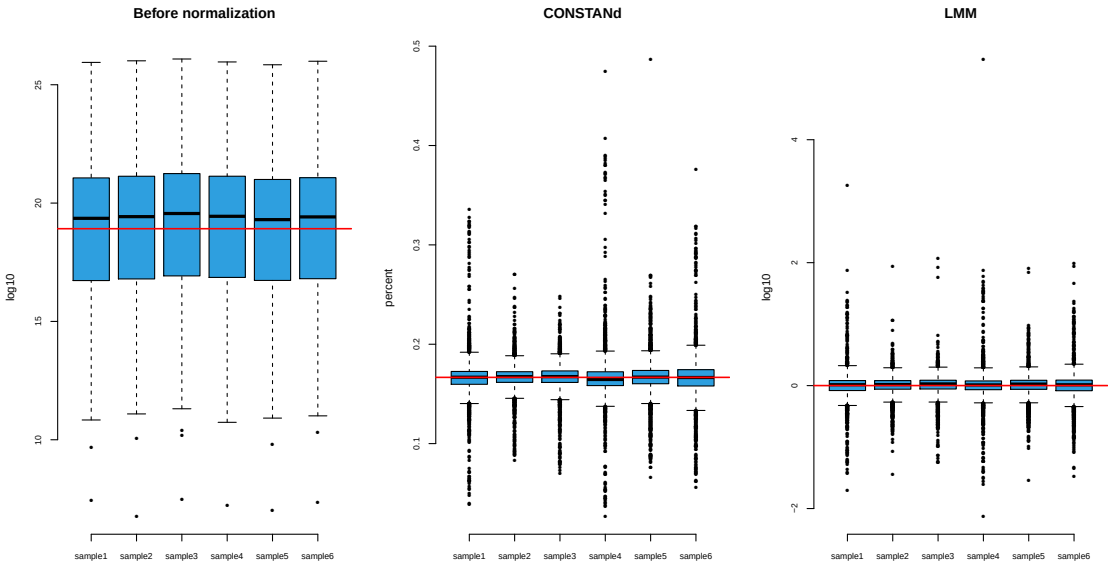

Figure 17: Boxplots of peptide reporter ion abundance from Erwinia dataset. The red line represents the mean value of peptide intensities across all MS^2 runs. Both normalization approaches result in similar boxplot profiles.

3.2.2 Protein fold change estimation

Since this is a spiked-in dataset where the true protein abundance is known, several variants of computing $\log_2$ FC may be considered, namely:

Approach 1 (A1): CONSTANd normalization followed by the $\log_2$ FC computation given by equation 5.

Approach 2 (A2): Model 9 normalization followed by the computation of $2^{w_{i,j(i),s,l}}$ used in the $\log_2$ FC estimation according to equation 5.

Approach 3 (A3): Model 9 normalization followed by the mean aggregation of the peptide repeated measurements in the normalized data. Next, a protein-by-protein linear modeling with the quantification channel as the only explanatory variable (fixed effect).

Approach 4 (A4): Model 9 normalization followed by the protein-by-protein linear modeling with the quantification channel as the only explanatory variable (fixed effect).

A comparison of A1 vs A2 may provide information on the effect of using CONSTANd normalized data (the peptide repeated measurements aggregated) data and the LMM normalized data (the peptide repeated measurements not aggregated) when simple $\log_2$ FC estimation is applied.

A comparison of A3 vs A4 may provide information on the effect of using the LMM normalized data in which the peptide repeated measurements were aggregated or were not aggregated, when $\log_2$ FC estimates are model based.

A comparison of A2 vs A4 may provide information on the effect of $\log_2$ FC estimation procedure (equation 5 or model based) obtained from the LMM normalized data (the peptide repeated measurements not aggregated)

The results of $\log_2$ FC estimation are presented in Figure 18.

Regarding the A1 and A2 comparison, it seems that using the LMM normalized data with the repeated measurements results in somewhat more accurate point estimates, but wider confidence intervals. This behaviour can observed in the majority of proteins and samples.

Regarding the A3 and A4 comparison, the $\log_2$ FC estimates yielded by both approaches are alike, but confidence intervals of A4 are narrower. This is due to not aggregating peptide repeated measurements in A4, which increases sample size and results in narrower confidence intervals.

A2 and A4 approaches both assume unaggregated input data but they differ in $\log_2$ FC estimation. It can be noticed that A2 estimates are closer to the true abundance of all proteins.

As a final remark, fitting a big model to Erwinia data, similar to that suggested by Oberg (Model 1), was not computationally feasible on a personal computer.

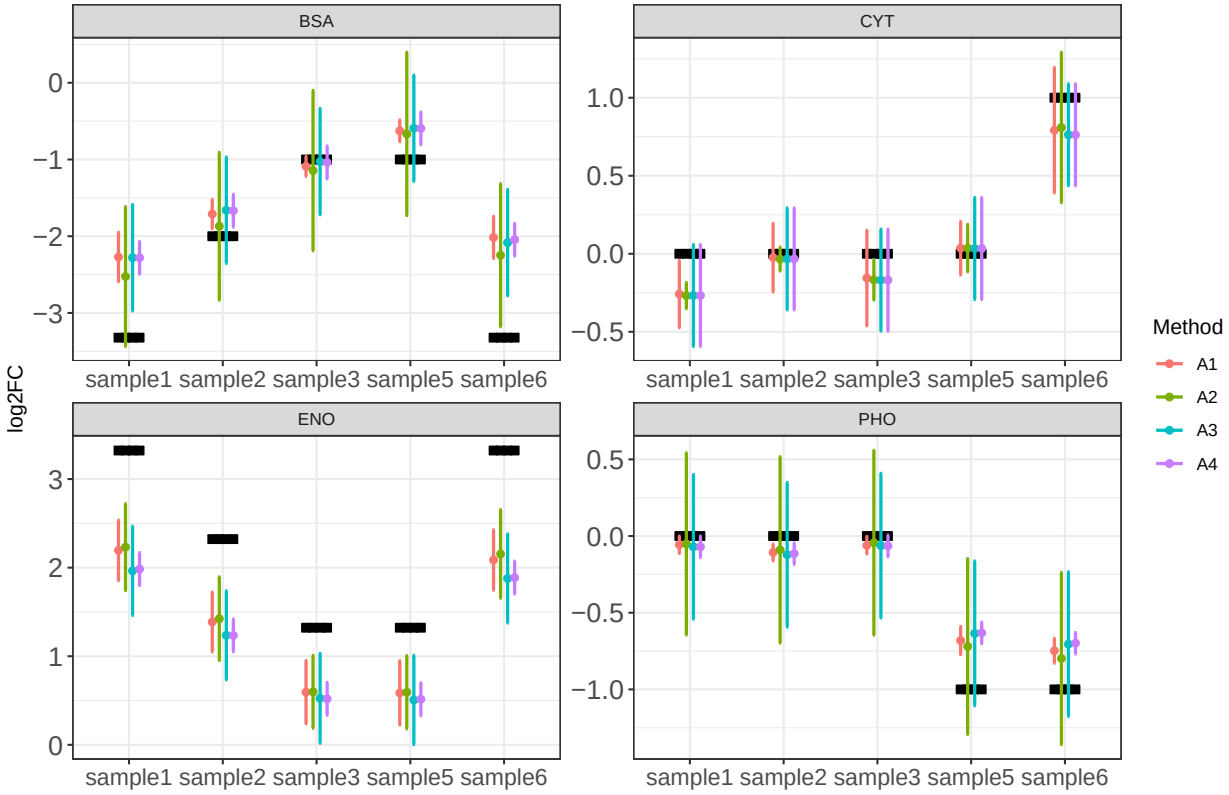

Figure 18: An overview of $\log_2$ FC estimation results.The panel names: BSA, CYT, ENO and PHO correspond to the four spiked-in proteins. The fourth sample was selected as the reference sample, therefore it is not shown on the graph. The black segments displays the true $\log_2$ FC protein fold changes. The coloured points represent $\log_2$ FC point estimates, while the vertical bars are confidence intervals of $\log_2$ FC estimates. Standard errors needed to compute the confidence intervals in A1 and A2 were obtained with the delta method.

4 Discussion

As it was emphasized in the beginning of this report, data normalization is a very important step before proceeding to differential expression analysis. Otherwise technical and experimental effects may mask the biological effects of interest. This was neatly illustrated in Table 2, which shows that the percentage of differential proteins based on unnormalized data is dramatically reduced as compared to analyses based on the normalized data.

This research has also shown that CONSTANd and linear mixed model normalization are both able to normalize peptide abundance data properly, as demonstrated in boxplots, PCA plots and heatmaps. However, the model-based normalization is a less automated process as compared to CONSTANd, because it requires additional attention with regard to several aspects.
First, one should decide on the type of an effect included in the model, namely, fixed versus random. It has already been mentioned that specifying the overall protein and peptide effects in the normalization model as random effects reduces computation time dramatically. The fixed effects specification would lead to approximately 10 hours of running time, or more, making the normalization procedure not very useful on a daily basis.
Second, in the case of MICE data, the normalization model (Model 2) proposed by Oberg et al. had to be adapted by adding the peptide-by-run random effect in order to remove between-run variation present on the peptide level. The peptide-by-run random effect resembles the row constraint in the CONSTANd method, which enables between-run comparisons. It can thus be hypothesised that every time a multiple MS^2 run study data is analyzed, the peptide-by-run random effect should be specified in the normalization model.

Besides removing unwanted technical effects, a data normalization method also affects differential expression analysis. From that point of view, CONSTANd and LMM based analysis results are somewhat different.
First, the spiked-in data results showed quite small differences in $\log_2$ FC estimates obtained from the CONSTANd based analysis (A1) and the LMM analyses (A2-A4). Similar differences can be spotted when looking at the MICE data results (the scatterplot on the right hand side of Figure 15). The differences observed on both datasets are relatively small, meaning that the peptide repeated measurements aggregation (not using all available information) have rather a little effect on $\log_2$ FC estimation. However, even small differences in $\log_2$ FC estimation may have some effect on p-values resulting from significance testing.

Significance testing could not be conducted on the Erwinia dataset because of too few samples analyzed in the study, but the MICE data showed important discrepancies in the percentage of differential proteins (Table 2) and the distribution of p-values (the scatterplot on the left hand side of Figure 15) obtained from CONSTANd based analysis and the LMM analysis. However, if the LMM analysis (data normalization and DEA) is repeated on the input data in which the peptide repeated measurements were aggregated, the percentage of differential proteins and the p-values distribution become more similar (see Table 3 and Figure 16).

Overall, the observed discrepancies in the significance testing results may arise due to:

1. Different input data. The LMM approach operates on full data hierarchy meaning that the peptide repeated measurements are not aggregated. On the other hand, in the CONSTANd based analysis the peptide repeated measurements are aggregated

prior to data normalization and then the normalized peptide quantification values corresponding to one protein are averaged within each sample. This results in only one quantification value per protein within a sample when performing a significance test. Hence two layers of the data hierarchy are summarized into one number and that may explain the discrepancies in p-values.

2. Different scale of peptide quantification values. The LMM based normalization and DEA assume $\log 2$ transformed data, while CONSTANd normalization convert the data to the [0,1] interval (percentage interpretation). It is know from probability theory that the expected value of a non-linear function of a random variable is not equivalent to the applying the non-linear function on the expected value of that random variable. The impact of different scales may be emphasised when the sample size is low, which is the case in protein-by-protein differential expression analysis.

3. Different normalization methods. It may just be that CONSTANd and linear mixed model both produce values that seems to be properly normalized (as presented on the boxplots, PCA plots and heatmaps included in this report), but there is still an underlying difference between the two normalization methods that affects further differential expression analysis. This aspect is the most interesting, but it needs to be disentangled from the two previous plausible sources of the observed discrepancies.

In summary, based on the conducted research, we cannot claim that the LMM based normalization and DEA outperforms (in terms of normalization quality, computation time, or DEA results) the CONSTANd normalization combined with ANOVA differential testing, or vice versa.

In order to unambiguously address these research questions, future work will include a similar comparison of the CONSTANd and LMM based analyses conducted on a bigger (more biological samples) spiked-in dataset or a simulated dataset. Hopefully, this will allow for deciding on the use of unaggregated versus aggregated data, choosing an appropriate scale for the differential expression analysis and assessing underlying differences in the CONSTANd and linear mixed model based normalization methods.
Data normalization and DEA in MS-based quantitative proteomics are popular research topics, therefore other normalization and differential testing procedures may be included in the aforementioned comparison.
Finally, it is conceptually attractive to perform data normalization and differential expression analysis in one model, therefore there will be an attempt to fit such a large model (Model 1) using a supercomputer or other high-computing environment.

References

[1] L. Martens I. Eidhammer, K. Flikka and S.-O. Mikalsen. *Computational Methods for Mass Spectrometry Proteomics*. John Wiley & Sons, Ltd, 2007.

[2] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology*, 2010.

[3] Qian Xu, Ziyou Cui, Gayathi Venkatraman, and Aldrin Gomes. The use of biophysical proteomic techniques in advancing our understanding of diseases. *Biophysical Reviews*, 4, 06 2012.

[4] Carl Murie, Brian Sandri, Timothy J. Griffin, Christine Wendt, and Ola Larsson. Normalization of mass spectrometry data (nomad). *bioRxiv*, 2017.

[5] Evelyne Maes, Wahyu Wijaya Hadiwikarta, Inge Mertens, Geert Baggerman, Jef Hooyberghs, and Dirk Valkenborg. Constand : A normalization method for isobaric labeled spectra by constrained optimization. *Molecular & Cellular Proteomics*, 15(8):2779–2790, 2016.

[6] Elizabeth G. Hill, John H. Schwacke, Susana Comte-Walters, Elizabeth H. Slate, Ann L. Oberg, Jeanette E. Eckel-Passow, Terry M. Therneau, and Kevin L. Schey. A statistical model for itraq data analysis. *Journal of Proteome Research*, 7(8):3091–3101, 2008. PMID: 18578521.

[7] Ann L. Oberg, Douglas W. Mahoney, Jeanette E. Eckel-Passow, Christopher J. Malone, Russell D. Wolfinger, Elizabeth G. Hill, Leslie T. Cooper, Oyere K. Onuma, Craig Spiro, Terry M. Therneau, and H. Robert Bergen, III. Statistical analysis of relative labeled mass spectrometry data from complex samples using anova. *Journal of Proteome Research*, 7(1):225–233, 2008. PMID: 18173221.

[8] Joris Van Houtven, Annelies Agten, Kurt Boonen, Geert Baggerman, Jef Hooyberghs, Kris Laukens, and Dirk Valkenborg. Qcquan: A web tool for the automated assessment of protein expression and data quality of labeled mass spectrometry experiments. *Journal of Proteome Research*, 18(5):2221–2227, 2019. PMID: 30942071.

[9] Laurent Gatto and Andy Christoforou. Using r and bioconductor for proteomics data analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(1, Part A):42 – 51, 2014. Computational Proteomics in the Post-Identification Era.